UNIVERSIDAD AUTÓNOMA DE MADRID

Escuela Politécnica Superior

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

ANÁLISIS DE DATOS PARA LA PREVENCIÓN DE FRAUDE EN APUESTAS DEPORTIVAS - DETECCIÓN DE AMAÑOS DE EVENTOS

Álvaro López de Prádena
Tutor: Celia Barbeyto Lanzas
Ponente: Ruth Cobos Pérez

Junio 2021

ANÁLISIS DE DATOS PARA LA PREVENCIÓN DE FRAUDE EN APUESTAS DEPORTIVAS - DETECCIÓN DE AMAÑOS DE EVENTOS

Autor: Álvaro López de Prádena
Tutor: Celia Barbeyto Lanzas

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Junio 2021

RESUMEN

El amaño de partidos es una práctica delictiva perseguida en todo el mundo por cada uno de los organismos que velan por la integridad en el deporte, es por ello por lo que en los últimos tiempos se está trabajando en herramientas para detectar y castigar a aquellos que participan en el fraude de las apuestas deportivas.

Por esta razón, en este Trabajo de Fin de Grado se trata de diseñar un mecanismo para analizar un importante número de partidos de tenis profesional, mediante el uso de técnicas como la detección de anomalías y modelos de *scoring*, para encontrar partidos en los que se haya podido cometer fraude.

Se dispone de un set de datos que contiene los resultados de los partidos de tenis pertenecientes tanto a la ATP como a la WTA de los últimos diez años (enero 2010 – octubre 2020), pero el diseño empleado permite ampliar el análisis a partidos de otros años y otras competiciones.

Para los encuentros analizados se crearon unas primeras variables en función del resultado del partido, el nivel de los jugadores (ranking) y sus cuotas en el partido. Para después incluirlas a la hora de realizar la detección de anomalías, donde se observaron los resultados para los algoritmos Isolation Forest y kNN, con distintos parámetros hasta conseguir el modelo con mejores resultados para distintas métricas de *clustering*.

Por último, calculado el riesgo de cada partido, se construyó un visualizador o *dashboard* con la tecnología Dash by Plotly en el que se pueden observar todos los detalles a nivel de evento y de jugador, para analizar su comportamiento y evaluar si efectivamente pudiera tratarse de un caso potencial de fraude; en el que además se incluyen los tweets escritos sobre los partidos más sospechosos, obtenidos a través de la API de Twitter.

Palabras clave Sistema Informático, Aprendizaje Automático, Detección de Anomalías, Visualización de Información.

ABSTRACT

Match fixing is an illegal practice prosecuted throughout the world by all of the organizations that ensure integrity in sport. This is why in recent times they are working on tools to detect and punish those who participate in sports betting fraud.

For this reason, this Bachelor Thesis tries to design a mechanism to analyze a significant number of professional tennis matches, using techniques such as anomaly detection and scoring models, to find matches in which fraud may have been committed.

There is a data set that contains the results of tennis matches belonging to both the ATP and the WTA from the last ten years (january 2010 – october 2020), but the design used allows the analysis to be extended to matches from other years and other competitions.

For each of the matches analyzed, first variables were created based on the result of the match, the level of the players (ranking) and their odds in the match. To later include them when performing anomaly detection, where the results for the Isolation Forest and kNN algorithms were observed, with different parameters until the model with the best results (for different clustering metrics) was achieved.

Finally, calculating the risk of each match, a visualizer or dashboard was built with Dash by Plotly technology in which all the details of the event and player level can be observed, to analyze their behaviour and evaluate whether it could indeed be a potential case of fraud; it also includes the tweets written about the most suspicious matches, obtained through the Twitter API.

Keywords Computer System, Machine Learning, Anomaly Detection, Information Display.

AGRADECIMIENTOS

Este trabajo no podría haber sido realizado sin los profesores y compañeros que me han acompañado durante mi paso por la universidad; con especial mención a mi tutora Celia Barbeyto por su ayuda con este trabajo, y también a mi ponente Ruth Cobos.

Además, quiero agradecer a mi familia que me ha dado fuerzas y ánimos en los peores momentos; además de ser mi principal apoyo en el día a día.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE CÓDIGOS

1 INTRODUCCIÓN

1.1 Motivación

El mundo de las apuestas deportivas es un universo que está en crecimiento en los últimos tiempos gracias a la aparición de las casas de apuestas online.

En España, en 2019 en el sector del juego se apostaron 35.665 millones de euros, produciendo unos ingresos de 10.226 millones de euros; mientras que sólo las apuestas deportivas supusieron ganancias de 394 millones de euros, con un nivel de juego de más de 2.000 millones de euros. El volumen de juego del sector y la relativa "facilidad" para amañar eventos deportivos, hace que las mafias se fijen en él [1].

Esta es una de las razones, por la que las organizaciones más importantes del mundo del deporte están dedicando grandes esfuerzos en mantener bajo control las apuestas que se realizan en sus competiciones, ya que se conocen varios casos recientes de amaños de partidos relacionados con las apuestas deportivas; que no hacen otra cosa que no sea ensuciar el deporte y la competición en la que son realizados.

Para frenar y mantener a raya este tipo de prácticas asegurando así la integridad de las competiciones, los organismos responsables, así como las casas de apuestas se están apoyando en el aprendizaje automático o *machine learning* y más concretamente en la detección de anomalías; para desarrollar herramientas que sean capaces de detectar eventos que hayan sido amañados. En este TFG se utiliza este tipo de herramientas para construir un software capaz de detectar posibles partidos sospechosos con datos que se encuentran de forma pública al alcance de cualquier persona.

1.2 Objetivos

Para alcanzar el objetivo principal de detectar partidos potencialmente sospechosos, se han de cumplir los siguientes objetivos:

- Construir una base de datos completa con los partidos de tenis disputados en los últimos diez años a máximo nivel internacional, esto son campeonatos ATP y WTA; todo ello con la máxima información a nivel de evento para poder realizar un análisis profundo.
- Definir y construir un modelo de datos específico de prevención de fraude, en el que se establezcan indicadores para los modelos de detección de anomalías y *scoring* de riesgo.
- Desarrollar una herramienta que nos permita obtener y almacenar la información de las redes sociales sobre cada uno de los partidos a analizar.
- Implementar un modelo de detección de fraude a partir del uso de la detección de anomalías entre otras técnicas.
- Crear un visualizador en el que se pueda observar toda la información disponible con claridad y gran nivel de detalle; además de disponer de una alta interactividad con el usuario.

Además, este Trabajo de Fin de Grado, implícitamente también tiene el objetivo de aprender nuevas tecnologías relacionadas con la detección de anomalías y la construcción de *dashboards*.

1.3 Estructura del documento

La memoria consta de los siguientes capítulos:

Capítulo 1. Introducción, se explican las motivaciones y los objetivos del documento.

Capítulo 2. Estado del arte, se presenta el estudio del estado del arte sobre los temas principales del trabajo.

Capítulo 3. Sistema desarrollado, se detalla el diseño y el desarrollo llevado a cabo en el proyecto.

Capítulo 4. Pruebas y resultados, se muestran los resultados obtenidos y las pruebas que han sido realizadas sobre el sistema.

Capítulo 5. Conclusiones y trabajo futuro, se dan las conclusiones obtenidas a la finalización del proyecto y el posible trabajo futuro que se podría hacer sobre el sistema.

Anexo. Visualización del sistema, se muestran las distintas vistas de las que dispone el visualizador construido.

2 ESTADO DEL ARTE

2.1 Apuestas deportivas

2.1.1 Introducción

La apuesta deportiva es el concurso de pronósticos sobre el resultado de uno o varios eventos deportivos, incluidos en los programas previamente establecidos por la entidad organizadora, o sobre hechos o actividades deportivas que formen parte o se desarrollen en el marco de tales eventos o competiciones por el operador del juego [1].

Esta actividad tiene su origen en las civilizaciones griegas y romanas cuando ya se hacían apuestas por ejemplo en los circos romanos por los gladiadores. Sin embargo, el crecimiento exponencial de esta práctica comenzó en el Reino Unido con las apuestas en las carreras de caballos y de galgos[3].

Además, una práctica que siempre han llevado de la mano las apuestas deportivas es el fraude deportivo, práctica en la que los apostadores amañan los eventos para conseguir el resultado que les interesa y así obtener grandes beneficios; para ello el método habitual es sobornar o amenazar a los protagonistas, ya sea algún jugador, entrenador e incluso los árbitros de la competición.

Para evitar este tipo de fraude, las federaciones hace tiempo que han puesto medidas como la prohibición de realizar apuestas a los participantes y hasta las personas de su entorno en algunos casos [3].

Más allá de los participantes directos en el acontecimiento deportivo sobre el que se realiza la apuesta, el BOE recoge la prohibición de apostar a estos otros grupos [1]:

- Los menores de edad.
- Las personas que han solicitado que se les prohíba el acceso el juego o lo tengan prohibido por una resolución judicial.
- Los accionistas o propietarios del operador del juego, así como su entorno cercano; ya sea de forma directa o indirecta a través de intermediarios.
- Los directivos de las entidades deportivas participantes u organizadoras del evento.
- Los jueces o árbitros del evento.
- Los miembros de la Comisión Nacional del Juego, así como su entorno más cercano.

2.1.2 Panorama Nacional

España es un país con una gran tradición deportiva, así como un país de apuestas como demuestra nuestra Lotería Nacional; uno de los juegos de lotería de mayor antigüedad en el mundo, cuyo primer sorteo se realizó en 1763.

Fue La Quiniela en 1946 la que puso en auge las apuestas deportivas como tal en nuestro país, un juego que aún hoy en día tiene bastante popularidad, aunque perdió muchos jugadores con la llegada de las apuestas deportivas online.

Fue en 2006, cuando las apuestas dejaron de estar únicamente en bares y casinos, para pasar a ser accesibles desde todas y cada una de las casas de nuestro país; todo ello gracias a las primeras leyes de apuestas en Internet aprobadas por el gobierno[4].

Desde entonces, el crecimiento de las casas de apuestas es imparable en España, como se puede ver con las cifras de las cantidades apostadas en juego online y concretamente en apuestas deportivas online (2013-2019) en las siguientes figuras; o sin irnos más allá como podemos observar en el día a día con la cantidad de establecimientos nuevos que se están abriendo, el número de anuncios que podemos ver en la televisión o el aumento de equipos de fútbol que tienen a casas de apuestas como principal sponsor [4][6].

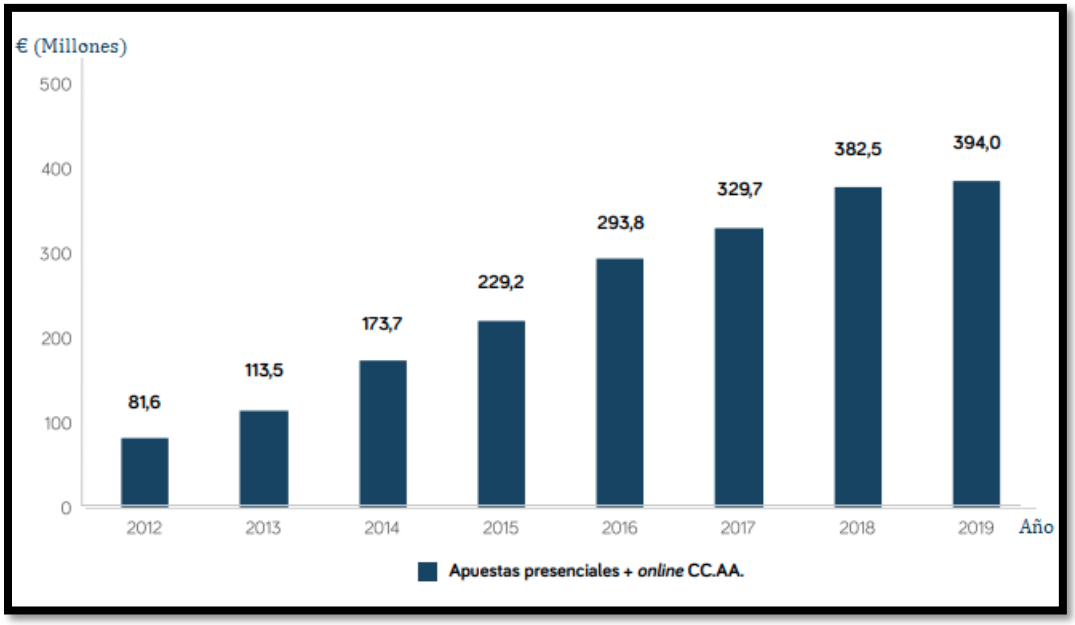


Figura 2.1: Juego real en apuestas deportivas en mercados regulados. Extraído de [1]. Crecimiento de las cantidades apostadas en juego presencial y online en España.

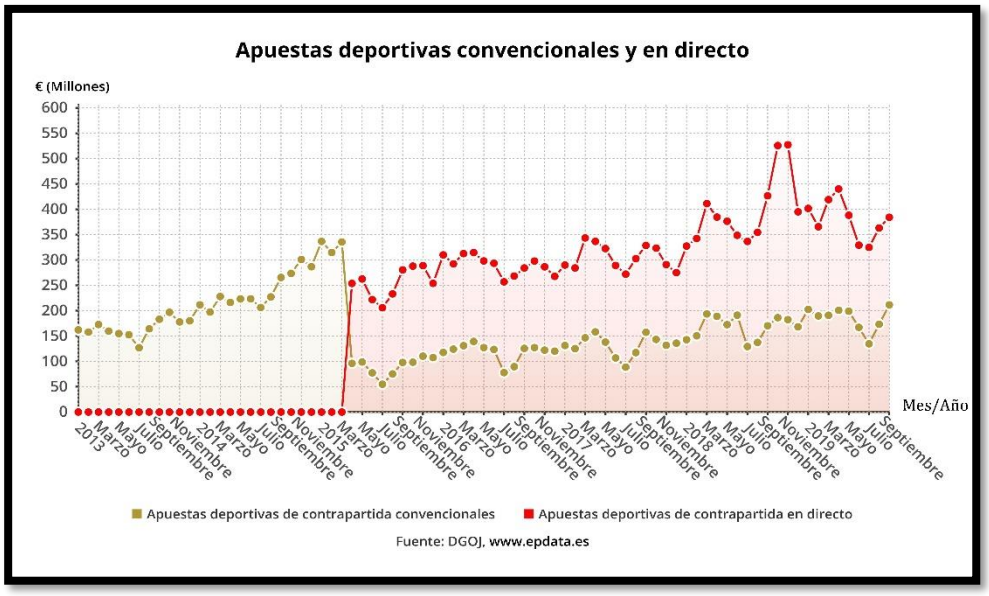


Figura 2.2: Juego en apuestas deportivas online de contrapartida. *Extraído de [4]*. Comparativa entre las cantidades movidas por las apuestas deportivas convencionales y las apuestas deportivas en directo.

Además, estas cifras son más preocupantes si cabe por el hecho de que el perfil de los jugadores es cada vez de personas más jóvenes, y por este motivo hay preocupación en el gobierno que en estos últimos meses ya ha empezado a tomar medidas como reducir la publicidad en televisión, así como limitarla a la franja horaria de la madrugada. Por otra parte, otra de las restricciones más importantes es la obligación de romper los contratos de patrocinio con las casas de apuestas; que afecta a los clubes de Primera y Segunda División de cara a la próxima temporada [7].

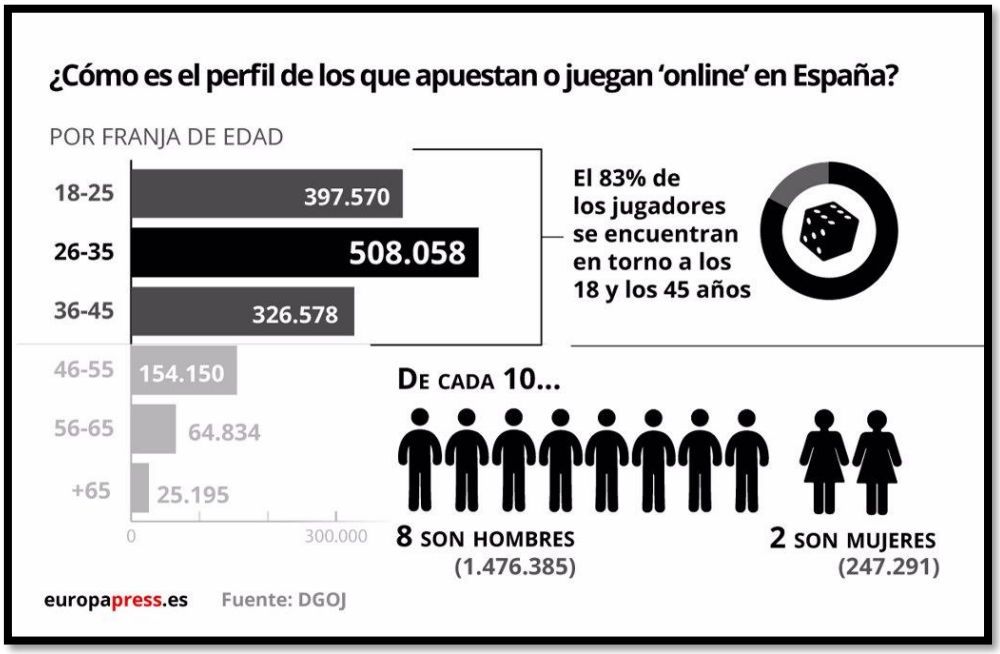


Figura 2.3: Perfil de los jugadores online en España. *Extraído de [4].*
Se muestran los perfiles de los jugadores de las apuestas en España por sexo y por edad.

2.1.3 Panorama en el Tenis

Las apuestas deportivas han llegado a prácticamente la totalidad de los deportes, pero como en todos los ámbitos, hay deportes con un porcentaje mayor de apuestas como pueden ser el fútbol por su gran número de aficionados o el boxeo y las carreras de caballos por su tradición en las apuestas.

Pero entre estos deportes que reciben un mayor porcentaje de apuestas también podemos encontrar al tenis, que además es uno de los deportes que más preocupa por el fraude deportivo en los últimos años; esto se debe en gran medida a que el tenis es un deporte individual fácilmente manipulable y por ello uno de los objetivos favoritos de las mafias que se encargan de organizar el fraude de las apuestas deportivas [8].

También, se sabe que el circuito ITF es el preferido entre los torneos del tenis, ya que en él participan los jugadores más desconocidos del circuito, con unos premios por victoria mucho menores que los de ATP por poner un ejemplo.

Esto implica, que para estos jugadores sea más sencillo caer en la tentación de un soborno o incluso arreglar partidos entre ambos jugadores [9].

Pero, asimismo se conocen numerosos casos de jugadores del circuito ATP que son sospechosos e incluso han sido sancionados por ser partícipes de actividades fraudulentas [10].

Por último, con el crecimiento de las redes sociales, en los últimos años cada vez hay más comentarios en estas sobre los partidos que facilitan encontrar comportamientos extraños; siendo mayor el número de comentarios a medida que la popularidad del encuentro es mayor, ya que partidos de menor categoría como los del circuito ITF son más complicados de seguir.

2.1.4 Tipsters Deportivos

Muchos de estos comentarios vienen de un nuevo perfil de usuarios que se ha hecho popular recientemente, estos son los llamados ‘tipsters’.

Son pronosticadores, supuestamente expertos en aquellos deportes en los cuales prometen a la gente grandes beneficios económicos si siguen sus consejos.

La mayoría de estas personas realmente cobran de las casas de apuestas por cada cliente que se crea una nueva cuenta, así como de los clientes que ingresan en sus canales privados (típicamente en Telegram).

Este realmente es su negocio y no los beneficios de las apuestas deportivas, pues trabajan prácticamente como captadores de las casas de apuestas [11].

2.1.5 Herramientas contra el fraude

Para luchar contra el fraude en las apuestas deportivas, los organismos que velan por el deporte están desarrollando herramientas capaces de detectar las prácticas fraudulentas para así poder castigarlas.

En esta situación, ha sido en España y más concretamente en ***LaLiga*** donde se ha desarrollado una de las herramientas antiamaños más potentes del mundo: Tyche 3.0, un software capaz de identificar cualquier movimiento extraño en los mercados de apuestas de fútbol y de detectar posibles fraudes en tiempo real.

El programa se encarga de monitorizar todos los encuentros aparecidos en más de cuarenta casas de apuestas que pertenecen desde a ***LaLiga Santander*** hasta al fútbol juvenil, pasando también por el fútbol femenino y el fútbol sala.

Para la detección de los partidos sospechosos utiliza *machine learning* basado en redes neuronales y tiene alertas para los cambios bruscos en cuotas, para las cuotas alejadas del modelo creado a través de la información disponible y una última con el partido en juego; en la que el programa hace saltar la alarma si el comportamiento en las casas de apuestas no concuerda con lo que ocurre en el terreno de juego [12].

Además, existen proveedores de soluciones tecnológicas como Sportradar o Iovation que están apostando por módulos específicos para la prevención del fraude en apuestas deportivas y en casino.

2.2 Redes sociales

2.2.1 Introducción

Con el auge de las redes sociales, es fácil encontrar información y opiniones de todo tipo en ellas, pues un gran porcentaje de la población tiene cuenta en al menos uno de los sitios web como Facebook, Instagram o Twitter.

Y gran parte del contenido que la gente sube a sus redes sociales se encuentra de forma pública y por tanto solemos disponer de un sencillo acceso a él.

Es por ello por lo que, a la hora de detectar partidos sospechosos de amaño, es interesante contar con los comentarios hechos por la gente sobre los partidos a analizar; puesto que más allá de los datos, de las redes sociales podemos recoger las sensaciones que han dejado los jugadores a los espectadores.

2.2.2 Twitter

Si hay una red social polémica y con grandes comunidades relacionadas con el deporte esa es Twitter, la red social del *hashtag* y del *retweet* es la preferida de la amplia mayoría de los seguidores del deporte. Y es donde más comentarios sobre cualquier evento deportivo podemos encontrar.

Además, esta red social cuenta con una práctica API que pese a las grandes restricciones con las que cuenta su uso gratuito (como veremos en el apartado de desarrollo), te permite realizar búsquedas y obtener datos de forma sencilla [13].

2.3 Detección de Anomalías

2.3.1 Introducción

La detección de anomalías es un problema usual de la ciencia de datos, cuyo objetivo es identificar observaciones, eventos o elementos extraños en los datos, que podrían ser indicativo de problemas en el proceso de recopilación de datos o de eventos inusuales como violaciones de seguridad. La detección de anomalías se puede llevar a cabo de forma supervisada, semi-supervisada o no supervisada [14][15].

Los **métodos supervisados** se dan cuando se conoce si cada observación, elemento o evento es anómalo o no; disponer de estas etiquetas para cada una de las observaciones es poco realista. Un ejemplo de estos métodos son los basados en algoritmos que utilizan una función de coste (*bagging* o *boosting*) [16].

Si disponemos de un conjunto de datos en el que conocemos las observaciones no anómalas, entonces deberemos emplear **métodos semi-supervisados**. Utilizando los datos no anómalos para el entrenamiento y evaluando si los datos no etiquetados son similares y se ajustan al modelo. Un ejemplo son los métodos basados en *kernel* como Máquinas de Soporte Vectoriales [17].

Cuando se sabe de la existencia de anomalías, pero no tenemos ninguna información etiquetada, se utilizan los **métodos no supervisados**; a este último tipo pertenecen los algoritmos probados en este trabajo (Isolation Forest y KNN).

2.3.2 Bosque de Aislamiento / Isolation Forest

Este algoritmo se basa en el funcionamiento de los Random Forest para determinar si un punto es anómalo o no. Para esto se crean distintos árboles que van haciendo divisiones sobre las distintas variables con el propósito de aislar las instancias del conjunto de datos; en función de la profundidad a la que quede aislado el punto, se le da una puntuación para poder determinar si es anómalo o no.

Los datos normales suelen llegar a mayores profundidades mientras que los anómalos quedan más cercanos a la raíz del árbol [18].

Al ser un método no supervisado, no existe el modo de conocer el valor óptimo desde el que se debe considerar una anomalía. La puntuación obtenida, es tan sólo una medida relativa respecto al resto de observaciones; es por ello por lo que suelen considerarse como potenciales valores atípicos, las observaciones cuya distancia está por debajo de un determinado cuantil [19].

Está considerado como uno de los mejores algoritmos de detección de anomalías por su eficacia y su sencilla implementación [20].

2.3.3 KNN

Este método suele estar basado en la asignación de la clase mayoritaria entre los k vecinos más cercanos a la instancia a clasificar. Sin embargo, al tratarse en este caso de un problema no supervisado, el proceso es semejante a crear *clusters*: se da por supuesto que las instancias normales tendrán una mayor similitud y cercanía, que las anómalas por tanto se distinguen dos *clusters* en los que un conjunto de instancias está diferenciado del conjunto ‘normal’.

El método puede utilizar un algoritmo por distancias como *K-Means* que busca el centro de un grupo de instancias, lo que conlleva que únicamente sea capaz de obtener *clusters* con formas esféricas; o también se puede usar un algoritmo como *DBScan* que se basa en crear *clusters* en función de la densidad de puntos cercanos [14].

2.3.4 Aplicaciones

Es un método que se suele emplear para detectar sucesos extraños, es decir, eventos cuyo comportamiento se salga del considerado como normal dentro de su conjunto; por tanto, algunas de las aplicaciones más usuales que se le da a la detección de anomalías son:

- Detectar intrusiones en la red
- Descubrir fraudes fiscales o en tarjetas de crédito
- Encontrar brotes de epidemias

2.3.5 Software

Existen diversas herramientas para utilizar el aprendizaje automático y en concreto la detección de anomalías, algunas de las más utilizadas son las siguientes:

- **Lenguaje de programación R**, lenguaje diseñado para el análisis estadístico. Su entorno de programación recibe el mismo nombre y es libre, es uno de los lenguajes de programación más empleados en la investigación; en los campos de aprendizaje automático o de minería de datos, por ejemplo.
- **La librería 'scikit-learn' de Python** [21], un conjunto de herramientas de Python para detectar anomalías en conjuntos de datos; incluye más de 30 algoritmos distintos.
- **Matlab**, herramienta de software matemático licenciado con un lenguaje de programación propio llamado M, es un programa muy utilizado en universidades y centros de investigación.

2.4 Visualizadores

2.4.1 Introducción

Un visualizador o *dashboard*, es una herramienta que nos permite mostrar nuestros datos de forma sencilla y óptima a través de elementos visuales como gráficas.

En los últimos años han surgido muchas herramientas de este tipo, cuyo uso a nivel empresarial y universitario está creciendo sobre todo a raíz del boom del *big data*.

Existen muchos servicios para el análisis de datos, todos ellos con compatibilidad con diferentes fuentes de datos, enormes posibilidades de personalización y grandes comunidades de usuarios que las utilizan. Algunos de ellos son:

- **Power BI,** la herramienta de Microsoft es una de las tecnologías más usadas en el sector de la visualización. Está más pensada en usuarios no familiarizados con el análisis de datos, lo que permite sacarle provecho desde el primer momento.
- **Tableau,** otra de las tecnologías líderes del sector, en este caso enfocada a un análisis más profundo y complejo; y cuenta con una enorme comunidad con una gran cantidad de recursos.
- **Grafana,** es una herramienta más enfocada al monitoreo de datos en tiempo real; requiere de mayores conocimientos que Power BI para su uso.
- **Dash by Plotly** [22]**,** desarrollado a continuación.

2.4.2 Dash by Plotly

Dash es un *framework* de Python pensado para construir aplicaciones web, que se utiliza frecuentemente para crear *dashboards* ya que permite una gran customización.

Basado en Flask, Plotly y ReactJ, no requiere de demasiados conocimientos para crear una sencilla aplicación web pues no es necesario saber HTML y Java Script para ello, al generarse completamente desde Python.

Por otra parte, al ser de código abierto existe la posibilidad de crear componentes propios como los que ha creado la comunidad para poder utilizar Bootstrap; y gracias a esto, cada día es más sencillo y hay más posibilidades a la hora de crear un visualizador interactivo utilizando Python.

3 SISTEMA DESARROLLADO

Arquitectura del sistema

La arquitectura del *Software* es de tres capas: datos, negocio y presentación. La primera se encarga de almacenar los datos de los partidos y los comentarios de las redes sociales. La capa de negocio, se ocupa de recibir las peticiones del usuario (comunicación con capa de presentación) y de responder tras procesar la petición (comunicación con capa de datos); en esta capa se incluye el módulo de procesamiento de datos (3.4). Finalmente, la capa de presentación se encarga de mostrar la información al usuario.

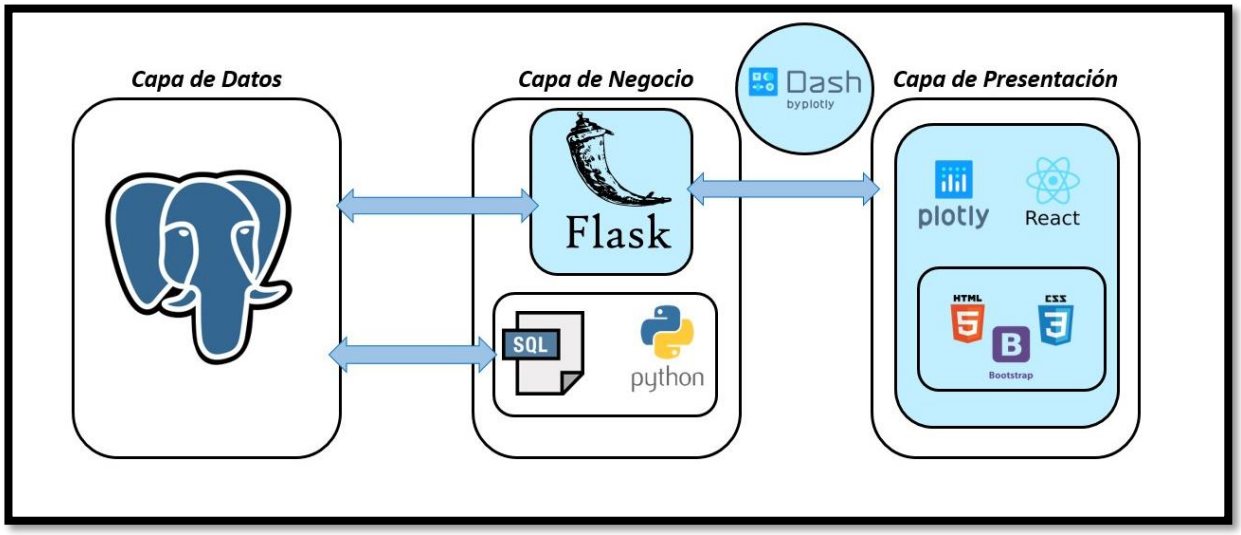


Figura 3.1: Diagrama de arquitectura.
Se muestran las distintas capas y las distintas tecnologías que forman parte del sistema

Además, en esta sección, en la Figura 3.2 se pueden observar los módulos que componen el sistema.

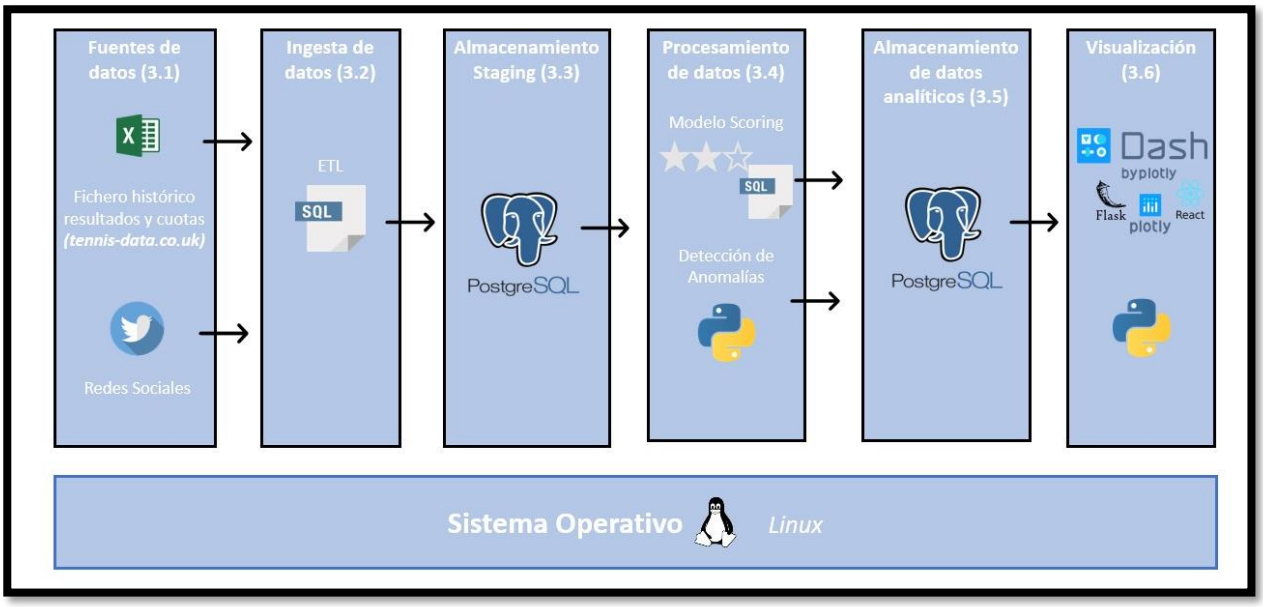


Figura 3.2: Diagrama funcional con los módulos que componen el sistema.
Se muestran los distintos módulos y tecnologías que componen el sistema.

En los siguientes apartados, se pueden ver los distintos módulos del proyecto, estos son:

- **Las fuentes de datos,** se exponen las fuentes de datos del proyecto.
- **La ingesta de datos,** se explica cómo se procesan los ficheros de resultados y cuotas, y cómo se extrae la información de Twitter.
- **El almacenamiento *staging*,** se explica dónde y cómo se guardan los datos de *staging*.
- **El procesamiento de datos,** dentro de este proceso se incluye la creación de atributos por una parte para la producción del modelo de *scoring* y por otra parte para su uso en la detección de anomalías; así como la detección de anomalías.
- **El almacenamiento de datos analíticos,** cómo están guardados los datos obtenidos del análisis.
- **La construcción del visualizador,** la herramienta de visualización construida y la forma de utilizarla para sacarla el máximo provecho.

3.1 Fuentes de datos

Los datos que contiene la base de datos provienen en su gran mayoría de la página web ***tennis-data.co.uk [23]*** y de la red social Twitter [13], todos ellos accesibles de forma pública. En esta base de datos se pueden encontrar hasta 53.054 partidos, pertenecientes a los diez últimos años (enero 2010 – octubre 2020) de las competiciones ATP y WTA; con toda la información relativa a ellos como la fecha del encuentro, los jugadores participantes, el resultado o las cuotas de las casas de apuestas (Bet365 [24] y Pinnacle [25]) por la victoria de cada jugador.

En la Tabla 3.1 se describen todas las columnas que forman parte de los archivos de datos utilizados.

Columna	Descripción
ATP/WTA	Código del torneo al que pertenece el partido en la asociación (ATP o WTA)
Location	Lugar donde se disputa el torneo
Tournament	Nombre del torneo al que pertenece el partido
Date	Fecha del partido
Series	Categoría del torneo
Court	Indica si el partido se juega en una pista interior o exterior
Surface	La superficie de la pista donde se juega el partido
Round	Ronda del torneo a la que pertenece el partido
Best of	El número de sets a los que se juega el partido
Winner	Jugador ganador del partido
Loser	Jugador perdedor del partido
WRank	Ranking del ganador del partido
LRank	Ranking del perdedor del partido
WPts	Puntos del ganador del partido
LPts	Puntos del perdedor del partido
W1	Juegos ganados en el primer set por el ganador del partido
L1	Juegos ganados en el primer set por el perdedor del partido
W2	Juegos ganados en el segundo set por el ganador del partido
L2	Juegos ganados en el segundo set por el perdedor del partido
W3	Juegos ganados en el tercer set por el ganador del partido
L3	Juegos ganados en el tercer set por el perdedor del partido
W4	Juegos ganados en el cuarto set por el ganador del partido
L4	Juegos ganados en el cuarto set por el perdedor del partido
W5	Juegos ganados en el quinto set por el ganador del partido
L5	Juegos ganados en el quinto set por el perdedor del partido
Wsets	Sets ganados por el ganador del partido.
Lsets	Sets ganados por el perdedor del partido.
Comment	Comentario sobre la finalización del partido. Ej. Suspendido.
B365W	Cuota del ganador del partido en Bet365.
B365L	Cuota del perdedor del partido en Bet365.
PSW	Cuota del ganador del partido en Pinnacle.
PSL	Cuota del perdedor del partido en Pinnacle.
MaxW	Cuota máxima del ganador del partido.
MaxL	Cuota máxima del perdedor del partido.
AvgW	Cuota media del ganador del partido.
AvgL	Cuota media del perdedor del partido.

Tabla 3.1: Columnas de los ficheros de datos.
Descripción de las variables que incluyen los ficheros de datos utilizados para crear la base de datos.

3.2 Ingesta de datos

3.2.1 Ficheros excel

Se utiliza un script que carga todos los datos de los resultados y cuotas de los partidos en la base de datos, y a continuación los formatea, modificando datos como los valores nulos, las fechas o el tipo de dato de las cuotas, así como se comprueba que no existen valores incorrectos como pueden ser los números negativos o las cuotas de valor menor a uno. Un fragmento de este script se muestra en el Código 3.1.

```
1.  UPDATE tenis_historico SET codigo_evento = CASE WHEN atp IS NOT NULL THEN atp ELSE wta END;
2.  UPDATE tenis_historico SET asociacion_id = CASE WHEN atp IS NOT NULL THEN 1 ELSE 2 END;
3.  ALTER TABLE tenis_historico ALTER COLUMN asociacion_id SET NOT NULL;
4.  ALTER TABLE tenis_historico DROP COLUMN atp;
5.  ALTER TABLE tenis_historico DROP COLUMN wta;

6.  UPDATE tenis_historico SET tier_id = CASE
7.  WHEN asociacion_id = 1 AND tier = 'Grand Slam' THEN 1
8.  WHEN tier = 'Masters 1000' THEN 2
9.  WHEN tier = 'Masters Cup' THEN 3
10. WHEN tier = 'ATP500' THEN 4
11. WHEN tier = 'ATP250' THEN 5
12. WHEN asociacion_id = 2 AND tier = 'Grand Slam' THEN 6
13. WHEN tier = 'Tour Championships' or tier = 'Elite Trophy' or tier = 'SEC' THEN 7
14. WHEN tier = 'Premier' THEN 8
15. WHEN tier = 'International' THEN 9
16. ELSE null END;
17. ALTER TABLE tenis_historico ALTER COLUMN tier_id SET NOT NULL;
18. ALTER TABLE tenis_historico DROP COLUMN tier;

19. UPDATE tenis_historico SET winner_rank = null WHERE winner_rank = 'N/A';
20. UPDATE tenis_historico SET loser_rank = null WHERE loser_rank = 'N/A';
21. UPDATE tenis_historico SET winner_points = null WHERE winner_points = 'N/A';
22. UPDATE tenis_historico SET loser_points = null WHERE loser_points = 'N/A';
23. UPDATE tenis_historico SET bet365_winner = REPLACE(bet365_winner, ',', '.');
24. UPDATE tenis_historico SET bet365_loser = REPLACE(bet365_loser, ',', '.');
25. UPDATE tenis_historico SET pinnacle_winner = REPLACE(pinnacle_winner, ',', '.');
26. UPDATE tenis_historico SET pinnacle_loser = REPLACE(pinnacle_loser, ',', '.');
27. UPDATE tenis_historico SET max_winner = REPLACE(max_winner, ',', '.');
28. UPDATE tenis_historico SET max_loser = REPLACE(max_loser, ',', '.');
29. UPDATE tenis_historico SET avg_winner = REPLACE(avg_winner, ',', '.');
30. UPDATE tenis_historico SET avg_loser = REPLACE(avg_loser, ',', '.');

31. ALTER TABLE tenis_historico ALTER COLUMN date TYPE date USING TO_DATE (date,
'DD/MM/YYYY')
```

Código 3.1: Fragmento del script que formatea los datos en la base de datos.
Se muestra como se crean nuevas variables a partir de otras y como se reemplazan símbolos, entre otras operaciones.

Por otra parte, en este módulo también está incluida la extracción de datos de Twitter.

3.2.2 Twitter: Extracción de datos

Una de las fases más problemáticas del proyecto ha sido la extracción de comentarios de las redes sociales, pues pese a que la API de Twitter tiene una gran funcionalidad frente a las de otras redes sociales como Instagram o Facebook que tienen grandes limitaciones a la hora de realizar búsquedas; también tiene grandes restricciones en cuanto a su uso gratuito se refiere sobre todo en búsquedas de contenido histórico.

Esto no habría sido un problema con el planteamiento inicial del proyecto que contemplaba el análisis de partidos en tiempo real, pero la pandemia acontecida durante el desarrollo de este trabajo y la suspensión de partidos que esta acarreó; obligó al cambio de planteamiento y el paso al análisis de partidos jugados hasta diez años atrás, lo que obligaba a usar la *Search* API (Full-Archive) [13] que en su versión gratuita restringe las peticiones a cincuenta al mes, los tweets por respuesta a cien y los caracteres de las consultas a ciento veintiocho.

Estas limitaciones forzaron a cambiar el uso de los datos extraídos de Twitter que inicialmente iban a ser utilizados a la hora de detectar las anomalías, para convertirse en una especie de factor comprobante; ya que únicamente se han obtenido los tweets de los partidos sospechosos tanto por el método de detección de anomalías como por el score creado.

Para asegurarnos de que los comentarios obtenidos tengan realmente valor se ha creado una consulta que busca términos como fraude, amaño o sospechoso; encontrando así tweets que señalan comportamientos sospechosos para los espectadores en aquellos partidos que nuestro sistema ha identificado como los más sospechosos entre los analizados.

Para poder utilizar la API de Twitter, previamente se debe tener una cuenta de desarrollador, llevar a cabo el registro de una aplicación para obtener las credenciales que nos permiten realizar peticiones; además de la configuración de un entorno de desarrollo, en este caso es ‘Search Tweets: Full Archive / Sandbox’ que es al que se puede acceder sin una suscripción de pago.

La petición se lleva a cabo en un programa realizado en Python, que guarda la respuesta en un JSON del cual se extrae la información recibida para ser almacenada en la base de datos a través de una conexión mediante la librería ***psycopg2 [26]***.

3.3 Almacenamiento *staging*

La base de datos relacional se crea en PostgreSQL utilizando el gestor pgAdmin 4, la elección de esta tecnología frente a otras alternativas como el servicio MySQL, se debe a la experiencia adquirida durante el grado.

En este módulo se tienen las siguientes tablas:

- **Tenis_historico,** la tabla principal en la que se insertan los datos a nivel de partido de las competiciones ATP y WTA. Se le insertan los datos extraídos de la fuente de datos, la web ***tennis-data.co.uk [23]*** de forma bruta; siendo formateados para adaptarse al diseño de la base de datos que cuenta con las tablas de dimensiones ya mencionadas.
- **Twitter_usuarios,** es una de las dos tablas que almacenan los datos extraídos de la red social Twitter; concretamente, esta tabla contiene la información sobre los usuarios que han realizado los tweets guardados en la siguiente tabla.
- **Twitter_historico,** la tabla que almacena toda la información sobre los tweets extraídos y los relaciona con el partido sobre el que se habla en el comentario.

3.4 Procesamiento de datos

El procesamiento de los datos incluye la generación de indicadores mediante el lenguaje SQL, así como la implementación del algoritmo de detección de anomalías.

A la hora de analizar el riesgo de cada evento se utilizan dos modelos: el modelo de *scoring* y el modelo de detección de anomalías a través de *machine learning*.

El modelo de scoring está formado por una serie de indicadores que dependen de las variables relevantes desde el punto de vista del negocio.

El modelo de detección de anomalías está formado por una serie de indicadores que enriquecen los datos sobre los que se lleva a cabo la detección de anomalías.

En los siguientes subapartados se describen ambos modelos y como se realiza la detección de anomalías.

3.4.1 Modelo de *scoring* de riesgo

Los indicadores creados suman puntos de riesgo a cada evento, cuya puntuación final se recoge en el score.

Este score ayudará a determinar los partidos que son sospechosos de fraude.

Los indicadores son iguales tanto en la tabla ATP como en la tabla WTA:

Nombre	Definición
flag_1	Se activa siempre y cuando el ganador sea el de la cuota mayor (menos favorito en las apuestas) y la diferencia entre las cuotas sea mayor que uno. Obteniendo como valor la mayor diferencia entre la cuota del ganador y el perdedor del partido.
flag_2	Se activa cuando las cuotas favorecen a un jugador en una casa de apuestas y a otro en otra de las casas de apuestas.
flag_3	Se activa si el partido se acaba por abandono de uno de los jugadores.
flag_4	Se activa si el ganador está cincuenta puestos o más por encima del perdedor (a mayor puesto en el ranking, peor clasificado) y además el ganador (peor por ranking) era el favorito de las apuestas.
score	Es la suma de las cuatro variables anteriores dándole un multiplicador por 2 a flag_2 y un multiplicador por 1.5 a flag_3.
sospechoso	Se activa cuando el score del evento es mayor a diez o si el evento es detectado como anomalía.

Tabla 3.2: Indicadores del modelo de *scoring* de riesgo.
Se definen los indicadores creados para el modelo de scoring.

3.4.2 Modelo de detección de anomalías

Se crean una sucesión de indicadores para llevar a cabo el análisis de los eventos con los algoritmos de aprendizaje automático para la detección de anomalías.

Coinciden los indicadores creados para las tablas de ATP y de WTA:

Nombre	Definición
ganador_promedio_ partidos_contraapuestas	Es el promedio (entre 0 y 1) de partidos del ganador del partido que son ganados pese a no ser el favorito de las apuestas.
perdedor_promedio_ partidos_contraapuestas	Es el promedio (entre 0 y 1) de partidos del perdedor del partido que son ganados pese a no ser el favorito de las apuestas.
jugadores_promedio_ partidos_contraapuestas	Es el promedio (entre 0 y 1) de partidos de los jugadores del partido que son ganados pese a no ser los favoritos de las apuestas.
ganador_promedio_ partidos_contraranking	Es el promedio (entre 0 y 1) de partidos del ganador del partido que son ganados pese a estar 20 puestos o más, peor clasificado en el ranking.
perdedor_promedio_ partidos_contraranking	Es el promedio (entre 0 y 1) de partidos del perdedor del partido que son ganados pese a estar 20 puestos o más, peor clasificado en el ranking.
jugadores_promedio_ partidos_contraranking	Es el promedio (entre 0 y 1) de partidos de los jugadores del partido que son ganados pese a estar 20 puestos o más, peor clasificados en el ranking.
ganador_promedio_ partidos_sospechosos	Es el promedio (entre 0 y 1) de partidos del ganador del partido que han activado alguna de las ‘flag’ creadas.
perdedor_promedio_ partidos_sospechosos	Es el promedio (entre 0 y 1) de partidos del perdedor del partido que han activado alguna de las ‘flag’ creadas.
jugadores_promedio_ partidos_sospechosos	Es el promedio (entre 0 y 1) de partidos de los jugadores del partido que han activado alguna de las ‘flag’ creadas.
ganador_promedio_ score	Es el promedio de score en los partidos del ganador del partido.
perdedor_promedio_ score	Es el promedio de score en los partidos del perdedor del partido.
jugadores_promedio_ score	Es el promedio de score en los partidos de los jugadores del partido.

Tabla 3.3: Indicadores del modelo de detección de anomalías.
Se definen los indicadores creados para el modelo de detección de anomalías.

3.4.3 Detección de anomalías

A la hora de llevar a cabo la detección de anomalías, el primer paso es **cargar los datos**, para ello se utiliza la librería ***psycopg2 [26]*** que nos facilita la conexión con nuestra base de datos; así como la librería ***pandas [27]*** que transforma los datos cargados en un *DataFrame.*

El segundo paso, es uno de los más importantes y el más laborioso, consiste en realizar **la preparación y la limpieza del conjunto de datos**.

Lo primero es asegurarse de que las columnas son de tipo numérico, para ello se le da a la columna *'Comment'*, que nos da información de cómo termina el partido, un valor numérico; así como se eliminan las columnas con los nombres de los jugadores.

Por otra parte, se deben eliminar los registros nulos y los registros duplicados, respecto a los duplicados ya han sido eliminados de la base de datos dentro del proceso de formateo. Para evitar los valores nulos, se usa la función *'isnull().sum()'* que muestra la suma de registros con valores nulos por columna; lo que nos ayuda a decidir qué columnas necesitan ser modificadas. Tras esto, se opta por sustituir los valores nulos por 0 en aquellas columnas en las que tiene sentido como las que dan la información sobre los juegos o sets de un partido; mientras que otros registros como los que tienen las columnas con la información de las casas de apuestas deben ser eliminadas, ya que son cuotas desconocidas y las cuotas no pueden ser 0.

Una vez preparado el conjunto de datos, el siguiente paso es **dividir el *DataFrame* en el conjunto de entrenamiento y el conjunto de *test*.**

Ya disponiendo de ambos conjuntos, **se construyen y se ajustan los modelos** (probamos con *KNN* e *Isolation Forest*) sobre el conjunto de entrenamiento utilizando la librería de Python ***sklearn [21]*** con la función *fit.* Para crear el modelo en esta librería, se utilizan parámetros como ***contamination*** o ***n_estimators*** (*Isolation Forest*). Para definir el primero de ellos (determina la proporción de anomalías en los datos), además de hacer pruebas se hace un estudio del valor más coherente con los datos analizados.

El entrenamiento, junto a la posterior predicción con el conjunto de *test* y la validación del modelo se llevan a cabo utilizando las siguientes funciones:

- *GridSearchCV*: Búsqueda exhaustiva de los mejores parámetros dado un diccionario con los parámetros y sus posibles valores; así como la forma de evaluarlos. En este caso se le dan las métricas de *clustering* utilizadas para evaluar el modelo, que se comentan más adelante. Además, este método hace uso de la validación cruzada y divide el conjunto de datos en k pliegues (siendo k un parámetro dado).
- *Silhouette score*: La primera y la más relevante de las métricas de *clustering* utilizadas para evaluar el modelo. Calcula el Coeficiente de Silhouette medio de las muestras, siendo el coeficiente la distancia entre una muestra y el *cluster* (siendo los *clusters* en este caso: 'Anómalías' y 'No anomalías') más cercano del que dicha muestra no forma parte [21].

- Davies Bouldin *score*: Segunda métrica de *clustering*. Se define como la medida de similitud promedio de cada grupo con su grupo más similar, siendo la similitud la relación entre las distancias dentro del grupo y las distancias entre grupos. A medida que los grupos estén más separados darán una mejor puntuación [21].
- Calinski Harabasz score: Última métrica utilizada. Se define como la relación entre la dispersión dentro de los *clusters* y la dispersión entre los *clusters [21]*.

Gracias a estas funciones, obtenemos el mejor modelo utilizando el algoritmo de Bosque de Aislamiento con los siguientes valores:

- Para los partidos ATP: 0.003 *contamination*, 10 *max_samples* y 70 *n_estimators* con 0.5 de *Silhouette Score.*
- Para los partidos WTA: 0.003 *contamination*, 70 *max_samples* y 10 *n_estimators* con 0.35 de *Silhouette Score.*

Durante la **fase de construcción y ajuste del modelo**, además de descartar el algoritmo KNN ya que se observa que tiende a distinguir los partidos de los mejores jugadores (*ranking* alto) como anómalos; también se descartan los partidos con comentario *'Walkover'* que tendían a ser distinguidos como anómalos al ser partidos que no llegan a ser disputados (por retirada o expulsión de alguno de los jugadores en la mayoría de los casos), pero como las apuestas de estos partidos son anuladas no tiene sentido alguno incluirlos en este análisis que busca los partidos sospechosos de fraude.

Respecto a la **validación del modelo**, al ser un problema no supervisado no se tiene una validación como tal, ya que no se puede disponer de un conjunto de datos etiquetados al no ser los partidos fraudulentos de dominio público; por tanto, se utilizan como método de validación las métricas de *clustering* que nos dicen cuál es el modelo que mejor está distinguiendo las muestras anómalas de las no anómalas; teniendo en cuenta para ello principalmente el *Silhouette Score* ya explicado.

Respecto al algoritmo elegido, el algoritmo *Isolation Forest*, consiste en aislar las observaciones seleccionando aleatoriamente una variable y escogiendo un valor aleatorio entre el máximo y el mínimo de dicha variable; así se va ramificando el árbol mientras que el valor de este registro sea mayor o menor que el valor aleatorio, acotando poco a poco al ser este valor aleatorio el nuevo máximo o mínimo. De esta forma, cuantas menos ramas necesite el árbol para aislar el punto, más anómalo será este. Finalmente, se calcula la media de todos los árboles (uno por cada variable) de cada instancia para conocer su grado de anomalía [20]. El método de la librería *sklearn* devuelve el conjunto de datos etiquetados, así como las puntuaciones de anomalía.

Una vez obtenidos los resultados con el modelo optimizado, son almacenados aprovechando la conexión ya creada con la base de datos.

3.5 Almacenamiento de datos analíticos

Además de las tablas de almacenamiento de staging vistas en el capítulo 3.3, en nuestra base de datos tenemos las tablas donde guardamos los datos analíticos, como se puede ver en la Figura 3.3:

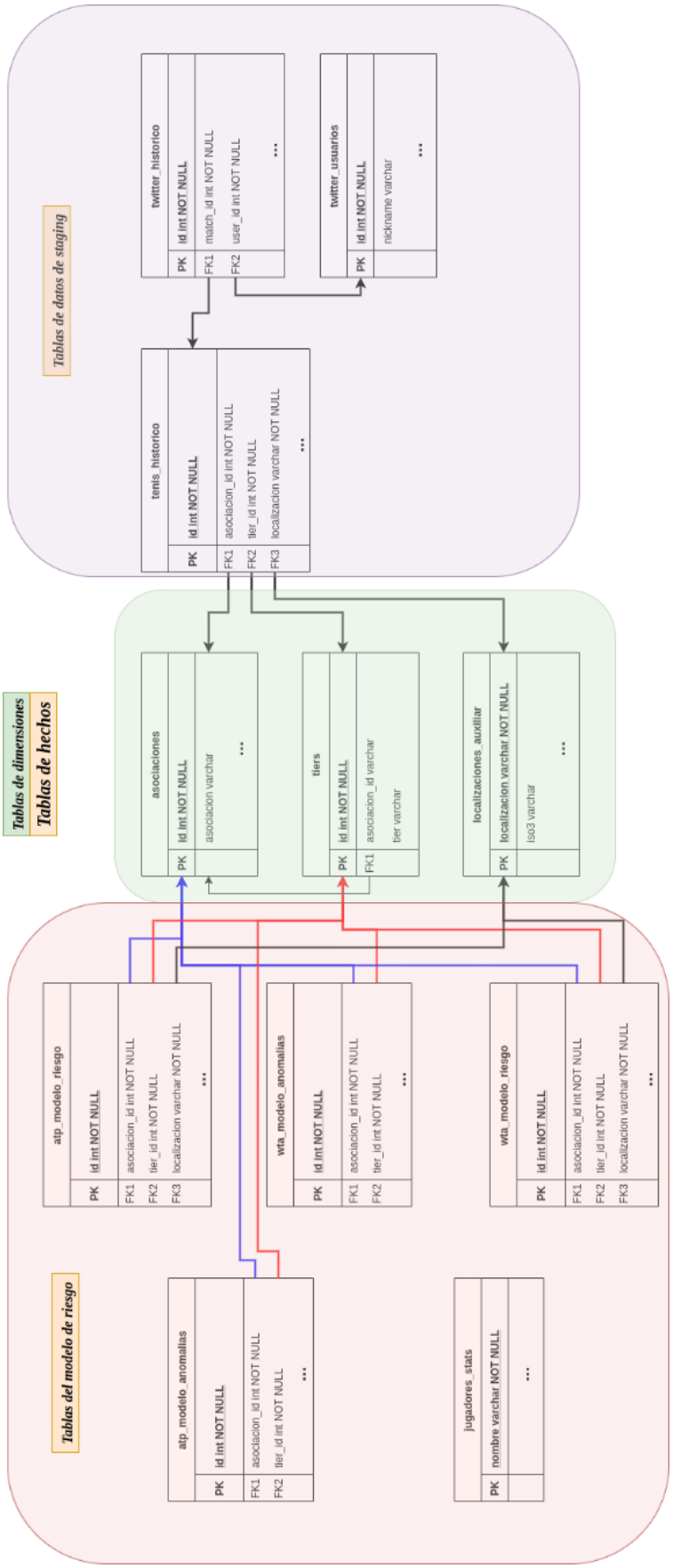


Figura 3.3: Diagrama relacional de la base de datos.
Se muestran las tablas de la base de datos y las relaciones existentes entre ellas

Tablas de dimensiones:

- **Asociaciones,** esta tabla contiene las distintas asociaciones que hay en la base de datos, estas son ATP y WTA en este estudio.
- **Tiers,** esta tabla contiene las diferentes categorías de torneos que existen en la base de datos.
- **Localizaciones_auxiliar,** como su nombre indica es una tabla auxiliar creada para el mapa creado en el visualizador; que contiene los códigos ISO [28] de los países en los que se realizan los partidos para poderlos situar geográficamente.

Tablas de hechos del modelo de datos analíticos:

- **ATP_modelo_riesgo,** esta tabla contiene los datos ya formateados de tenis_historico que pertenecen a la ATP y además guarda varias variables calculadas que se explican más adelante en este apartado.
- **WTA_modelo_riesgo,** esta tabla contiene los datos ya formateados de tenis_historico que pertenecen a la WTA y además guarda varias variables calculadas que se explican más adelante en este apartado.
- **ATP_modelo_anomalias,** esta tabla contiene todas las variables utilizadas en el algoritmo de detección de anomalías y los resultados del mismo para los eventos pertenecientes a la ATP.
- **WTA_modelo_anomalias,** esta tabla contiene todas las variables utilizadas en el algoritmo de detección de anomalías y los resultados de este para los eventos pertenecientes a la WTA.
- **Jugadores_stats,** esta tabla contiene datos a nivel de jugador como los partidos disputados o las victorias, y su creación está principalmente motivada por el visualizador.

3.6 Visualizador

El objetivo final de la herramienta creada es la visualización del análisis realizado, en una interfaz gráfica interactiva con el usuario.

Para ello se elige la tecnología Dash by Plotly [22] que nos permite crear un *dashboard* en Python, como hemos visto en el estado del arte, Dash es un *framework* dedicado a la construcción de aplicaciones web analíticas.

Escrito sobre Flask, Plotly.js y React.js nos permite desarrollar una aplicación de visualización de datos con una interfaz personalizada sin tener demasiados conocimientos de HTML o CSS, ya que se escribe de principio a fin en el lenguaje de programación Python.

La fuente de datos del *dashboard* es la base de datos creada, que se conecta al visualizador a través de la librería ***psycopg2 [26]*** ya mencionada en anteriores apartados.

Para dotar de estilo a la interfaz, se utiliza la librería ***dash_bootstrap_components [28]*** desarrollada por la comunidad de Dash que permite emplear temas de Bootstrap.

El visualizador está organizado en tres diferentes pestañas: General, Partidos Sospechosos y Buscador.

Las distintas vistas que tiene el visualizador se pueden observar en el Anexo A, donde se puede ver cómo al inicio se requiere el ingreso de un usuario y una contraseña para poder acceder al contenido (véase Figura A.1).

A continuación, se explican las distintas secciones que existen en el *dashboard* y el funcionamiento que tienen.

3.6.1 General

En esta pantalla principal, se ofrece una visión global de todos los datos que han sido sometidos al análisis (véase Figura A.2 y Figura A.3).

Para comenzar se dispone de una tarjeta con el número de partidos que se están visualizando en cada momento (aquellos que cumplen con los filtros).

Poco más abajo, nos encontramos con los filtros disponibles para todas las gráficas de esta sección, estos son: asociación, categoría, superficie, tipo de partido (sospechoso o todos) y rango de fechas.

A continuación, se pueden encontrar las siguientes gráficas de tarta:

- **El porcentaje de partidos por asociación.**
- **El porcentaje de partidos por categoría (*tier*).**
- **El porcentaje de partidos por superficie.**
- **El porcentaje de partidos sospechosos. ***

Y, por último, encontramos una gráfica de barras con el número de partidos (sospechosos y no sospechosos diferenciados) por cada mes incluido en el análisis.

De esta forma, podemos buscar por ejemplo el número de partidos sospechosos de fraude que hay en ATP, en torneos Grand Slam que se juegan en hierba, desde agosto de 2015.

*****Un ejemplo del código desarrollado para implementar concretamente la gráfica señalada, se muestra en Código 3.2 y Código 3.3.

```
32.     dcc.Graph(
33.         id = 'grafica_partidos_sospechosos',
34.         figure = px.pie(data.groupby('sospechoso')['id'].agg('count').to_frame(name =
'partidos').reset_index(),
35.                 values = "partidos",
36.                 names ="sospechoso",
37.                 title = '% Partidos Sospechosos',
38.                 template = "seaborn",
39.                 labels ={'sospechoso': 'Es Sospechoso', 'partidos': 'Partidos'})
40.     )
```

Código 3.2: Creación de una gráfica de tarta con Dash.
Se muestra el código utilizado para crear la gráfica de tarta utilizando la tecnología Dash by Plotly.

```
41.     @app.callback(
42.             Output('grafica_partidos_sospechosos', 'figure'),
43.             [Input('fechas1', 'start_date'), Input('fechas1', 'end_date'),
44.             Input('asociacion1', 'value'), Input('tier1', 'value'),
45.             Input('superficie1', 'value'), Input('estado1', 'value')])
46.     def update_grafica_partidos_sospechosos(start_date, end_date, asociacion, tier, surface,
tipo):
47.             if asociacion != 'Todas' and asociacion != None:
48.                     filter_data = data[(data['asociacion'] == asociacion)]
49.             else:
50.                     filter_data = data
51.             if tier != 'Todas' and tier != None:
52.                     filter_data = filter_data[(filter_data['tier'] == tier)]
53.             if surface != 'Todas' and surface != None:
54.                     filter_data = filter_data[(filter_data['surface'] == surface)]
55.             if tipo != 'Todos' and tipo != None:
56.                     filter_data = filter_data[(filter_data['sospechoso'] == True)]
57.             filter_data = filter_data[(filter_data['date'] >= start_date) & (filter_data['date'] <=
end_date)]
58.
59.             return px.pie(filter_data.groupby('sospechoso')['id'].agg('count').to_frame(name =
'partidos').reset_index(),
60.                     values = "partidos",
61.                     names ="sospechoso",
62.                     title = '% Partidos Sospechosos',
63.                     template = "seaborn",
64.                     labels ={'sospechoso': 'Es Sospechoso', 'partidos': 'Partidos'})
```

Código 3.3: Actualización de una gráfica en función de los filtros seleccionados.
Se muestra el código utilizado para actualizar la gráfica de tarta utilizando la tecnología Dash by Plotly.

3.6.2 Partidos Sospechosos

En esta pantalla del visualizador, el análisis se centra en los partidos sospechosos de fraude (véase Figura A.4 y Figura A.5).

En la parte superior se encuentra una gráfica de burbujas en la que se puede observar la disposición de los jugadores en función de los ejes: promedio de partidos anómalos y promedio de partidos de riesgo. Esto nos facilita encontrar a los jugadores más sospechosos de haber cometido fraude, aunque cuenta con el inconveniente de que los jugadores con menos partidos disputados son proclives a situarse en las partes más destacadas al tener mayor facilidad para tener promedios altos.

Al lado de esta gráfica encontramos un mapa con los partidos sospechosos que hay por país, en el que además se puede diferenciar por continente.

Debajo de estas gráficas, como en la pestaña general nos encontramos con un contador de los partidos que se están visualizando y los filtros que se pueden aplicar sobre las gráficas que se encuentran en la parte inferior; estos filtros son los mismos que se han descrito en la anterior sección a excepción del filtro de tipo de partido, ya que en esta pantalla todos los partidos son sospechosos.

Las gráficas incluidas en esta sección, además, incorporan la funcionalidad que permite ver la lista de partidos sospechosos al pinchar sobre un sector en las gráficas de tartas o sobre un mes en las gráficas de barras. Esto posibilita, que conociendo un partido luego puedas buscar más información sobre él en el buscador.

Son las siguientes:

- **Gráfica de tarta con el porcentaje de partidos por motivo del final de partido**, esto es si el partido fue completado, si hubo abandono, etc.
- **Gráfica de tarta con el porcentaje de partidos por el método de detección,** esto es si es sospechoso por haber sido detectado como anomalía, por tener un score de riesgo mayor a diez o por cumplir ambas condiciones.
- **Gráfica de barras con los partidos sospechosos por mes.**

3.6.3 Buscador

En esta última pestaña, podemos encontrar más información sobre los jugadores o sobre partidos concretos.

En su pantalla inicial, encontramos una barra de búsqueda que nos permite buscar al jugador en el que estemos interesados.

Una vez seleccionado el jugador, podemos elegir si ver información sobre el jugador o sobre alguno de sus partidos sospechosos en concreto (véase Figura A.6).

Si se opta por la primera opción, encontramos una serie de datos como su mejor y peor ranking, su número de partidos y sus victorias; así como comparativas entre sus promedios con los de la media y la media de su género, entre estos promedios podemos encontrar el de victorias o el de partidos sospechosos (véase Figura A.7).

Por otra parte, hay una parte inferior en la que encontramos filtros y gráficas como en las anteriores pantallas (véase Figura A.8 y Figura A.9); estas gráficas son:

- **Gráfica de tarta con el porcentaje de partidos por categoría.**
- **Gráfica de tarta con el porcentaje de partidos por superficie.**
- **Gráfica de tarta con el porcentaje de partidos por motivo del final del partido.**
- **Gráfica de tarta con el porcentaje de partidos sospechosos. ***
- **Gráfica de tarta con el porcentaje de partidos sospechosos por método de detección. ***
- **Gráfica de barras con el número de partidos por año (diferenciando sospechosos y no sospechosos). ***
- **Gráfica de barras con el score medio por mes.**

***** En las gráficas señaladas, está implementada la funcionalidad que permite ver la lista de partidos sospechosos al pinchar sobre ellos.

Si, por el contrario, se opta por elegir un partido sospechoso del jugador accedemos a un despegable que nos permite buscar y seleccionar el partido en el que haya interés.

Ya seleccionado el partido, se muestran todos los datos disponibles en la base de datos sobre el evento como la asociación, el torneo y su localización, la ronda, los jugadores, el resultado, etc. (véase Figura A.10).

Además, se puede ver si el partido ha sido detectado como anómalo y una serie de comparativas entre las cuotas del partido y las cuotas medias o el score frente al score medio (véase Figura A.11).

Por último, en aquellos partidos de los que se hayan extraído comentarios valiosos de Twitter, se podrá observar una tabla con los tweets y datos que indican su relevancia como los *retweets* obtenidos o el autor del comentario y sus seguidores (véase Figura A.).

4 PRUEBAS Y RESULTADOS

4.1 Pruebas

En este capítulo, se enumeran las pruebas realizadas para obtener los resultados deseados y descritos en el siguiente apartado.

Las pruebas se pueden agrupar en dos grandes bloques:

Visualizador

- Pruebas de integración del sistema.
  - Conexión entre el visualizador (Dash by Plotly) y la base de datos.
- Pruebas de caja negra.
  - Situaciones de excepción.
  - Valores límite.
  - Transiciones de estados.

Además, dentro de este mismo bloque de pruebas, también se incluye el ***test* de usabilidad** realizado con la técnica 'Thinking Aloud' cuyo uso se detalla en el anexo *Thinking aloud*.

Análisis de eventos anómalos

En este segundo bloque se incluyen las siguientes pruebas realizadas:

- **La detección de anomalías**.
  - Se evalúan (con las métricas ya mencionadas en su apartado) los resultados obtenidos para los algoritmos de detección de anomalías *KNN* e *Isolation Forest*; probando para cada uno de ellos distintos valores para sus parámetros.
  - Como se deben almacenar los resultados en la base de datos, también se comprueba la conexión con la misma.
- **La extracción de datos de Twitter a través de la API**.
  - Se corrobora que las peticiones realizadas a la API son respondidas hasta alcanzar el límite mensual.
  - Su correcto almacenaje dentro de la base de datos comprobando que la conexión con esta es correcta.

Por último, se procede con **la validación de los resultados obtenidos** y para ello se estudia una muestra aleatoria de los partidos que los dos modelos desarrollados coinciden en señalar como sospechosos de fraude, junto a una experta en el área de prevención de fraude.

* Por asunto de protección de datos no se puede dar la identidad de los jugadores partícipes en los encuentros estudiados.

En el **partido 1 (WTA International, septiembre 2012)**, se observan claramente comportamientos característicos de los partidos fraudulentos ya que el jugador 1 vence por abandono del jugador 2 en el primer set de la primera ronda del torneo; siendo el jugador 2 claramente favorito en las apuestas al estar entre los 20 mejores del ranking mientras que el jugador 1 tiene un ranking muy bajo; es por ello por lo que las cuotas al ganador del partido son altas manteniéndose en 26 euros por euro apostado en una de las casas de apuestas de las que se poseen datos.

Una vez detectados estos comportamientos sospechosos, se contrastaría el importe de apuestas al ganador (datos no públicos) para comprobar si realmente ha habido comportamientos fraudulentos alrededor de este partido.

En el **partido 2 (WTA International, julio 2011)**, también se dan situaciones típicas de partidos fraudulentos como el abandono del jugador favorito de las apuestas, en este caso en el segundo set; sin embargo, sobre este partido es conocido que el jugador se retiró por una lesión.

Aunque se dan las condiciones para ser considerado como sospechoso por los modelos desarrollados, se puede descartar la actividad fraudulenta en este partido con la información disponible.

En el **partido 3 (ATP Grand Slam, octubre 2020)**, se encuentra de nuevo un caso en el que el jugador 1 con un ranking muy bajo consigue la victoria ante uno de los 20 mejores clasificados del ranking, a diferencia de los anteriores partidos en este caso no hay una retirada del jugador 2. Sí que sigue el comportamiento habitual de tener una cuota muy superior el ganador y tratarse de una de las primeras rondas del torneo. Además, el jugador 1 tiene un promedio de partidos anómalos superior a la media.

Observadas estas sospechosas conductas, se debería comprobar el importe de apuestas al ganador (información privada de las casas de apuestas) para corroborar si ciertamente hubiera comportamientos fraudulentos relacionados con este partido.

El **partido 4 (ATP Masters 1000, octubre 2014)** es un caso representativo de partido sospechoso ya que es un partido de primera ronda en el que el jugador 1 es un invitado al ser un torneo de su país y tiene un ranking muy bajo, pero aún así es capaz de vencer al jugador 2 (ranking alto) en dos sets. Además, sobre este partido existen comentarios de las redes sociales que también señalan un extraño comportamiento del jugador 2 a lo largo del partido. Las cuotas del ganador son ampliamente superiores a las del perdedor en las casas de apuestas de las que se dispone información.

Analizadas las sospechosas prácticas de este partido señalado por los modelos desarrollados como sospechoso, se debería ver el importe de apuestas al ganador del partido (datos no disponibles de forma pública) para comprobar si realmente ha habido una actividad fraudulenta ligada a este evento.

En último lugar, el **partido 5 (WTA International, abril 2010)** tiene bastantes similitudes con el **partido 3,** estando mucho mejor posicionado en el ranking y siendo claramente favorito para las casas de apuestas el jugador derrotado. Además, esta derrota también se da en una de las primeras rondas del torneo y no hay una retirada del jugador derrotado.

Como en anteriores casos, habría que comprobar el importe de apuestas al vencedor (información privada) para poder afirmar que el partido esté relacionado con alguna actividad fraudulenta.

4.2 Resultados

Gracias a las pruebas realizadas, tenemos un *dashboard* completamente funcional y unos resultados validados que analizaremos en este apartado de forma global, así como de forma más concreta.

Las pruebas de integración del sistema, así como todas las pruebas de caja negra han resultado satisfactorias, y el visualizador en ningún momento da salidas inesperadas o erróneas; no aceptando la entrada de valores inválidos como podrían ser los nombres de jugadores inexistentes o la búsqueda de partidos de ATP de tier 'International' como ejemplos. La aplicación en todo momento funciona según lo esperado.

Por otra parte, se han obtenido unos resultados satisfactorios tras llevar a cabo la técnica *Thinking aloud.* A continuación, se muestra la retroalimentación obtenida:

- Mejores características:
  - Es una herramienta intuitiva para navegar.
  - Da información relevante para el análisis.

- Aspectos que podrían mejorarse:
  - Al usuario le cuesta volver a la vista original de las gráficas una vez han hecho *zoom*, debería haber una pequeña explicación de los iconos que incluye Plotly para interactuar con las gráficas.
  - Se debería mantener el criterio de color en las gráficas para partidos sospechosos y no sospechosos, ya que da lugar a alguna confusión.
  - Se propone la idea de navegar a la pantalla de detalle de un partido al pinchar sobre él, en lugar de tener que ir posteriormente a buscarlo en otra pestaña.
  - Sería adecuado tener una pequeña explicación de los criterios a la hora de calcular alguna métrica como puede ser la cuota media del partido.

En general, el usuario está contento con el funcionamiento de la aplicación y esto se ve reflejado en la encuesta recibiendo una media de 3.5 (entre 1 y 5) en las distintas calificaciones que se proponen en el formulario como se puede observar en la Figura B.1 y en la Figura B.2.

La característica con mayor margen de mejora, como ya se ha comentado entre los aspectos a mejorar, es la navegación entre pantallas que se considera que podría ser más rápida. Por último, las tareas del *test* de usabilidad han sido completadas en el tiempo estimado.

Así como, gracias a las pruebas del bloque de análisis de eventos anómalos:

- Se ha desarrollado un modelo de *machine learning* utilizando el algoritmo Isolation Forest.
- Se han almacenado los comentarios de la red social Twitter de muchos de los partidos analizados.
- Se ha validado una muestra representativa de los partidos detectados como sospechosos por los dos modelos desarrollados (*scoring* y detección de anomalías) en este trabajo.

A continuación, se pueden observar los resultados obtenidos tras llevar a cabo el análisis de los 53.054 eventos de la base de datos.

En la Tabla 4.1, se pueden ver los partidos señalados como sospechosos en función del método que les ha detectado como dudosos ya sea el modelo de *scoring* de riesgo, el modelo de detección de anomalías o ambos.

Estos eventos señalados como sospechosos no tienen por qué realmente haber sido fraudulentos, sino que el algoritmo de detección de anomalías o el modelo de *scoring* los han señalado por salirse del comportamiento considerado normal respecto al resto de los eventos desde un punto estadístico.

Como se ha comentado en la sección de pruebas, sería estrictamente necesario realizar una investigación profunda de cada caso con datos no disponibles de forma pública para poder concluir que realmente haya habido un comportamiento fraudulento.

Asociación	Número de Partidos Analizados	Número de Partidos Sospechosos	Número de Partidos Anómalos con Score <= 10	Número de Partidos No Anómalos con Score > 10	Número de Partidos Anómalos con Score > 10
ATP	27.301	146	78	65	3
WTA	25.753	103	67	28	8
Todas	53.054	249	145	93	11

Tabla 4.1: Resultados del análisis realizado, partidos sospechosos.
Se resumen los resultados obtenidos en el trabajo tras el análisis completado.

Con estos resultados, se puede calcular que el porcentaje de partidos sospechosos de amaño es cercano al 0.5%; siendo superior a este porcentaje si sólo tomamos los partidos ATP e inferior si por el contrario escogemos únicamente los de WTA.

Por otra parte, hay que tener en cuenta que en hasta 4 de los partidos (1 ATP y 3 WTA) señalados como sospechosos por el modelo de *scoring* se produjo el ***walkover***, esto es la renuncia o la expulsión de uno de los jugadores antes de comenzar el partido (en la mayoría de los casos por lesión).

El *walkover* anula las apuestas realizadas en ese partido, por lo que son partidos en los que la actividad fraudulenta no tiene razón de ser.

A continuación, se llevan a cabo dos análisis más minuciosos sobre los resultados obtenidos que aportan una información interesante.

Partidos sospechosos por superficie:

- **En tierra batida o arcilla**: hay **73** partidos sospechosos, un 29.3% del total.
- **En hierba**: hay **38** partidos sospechosos, un 15.3% del total.
- **En pista dura o cemento**: hay **138** partidos sospechosos, un 55.4% del total.

En la Figura 4.1 se muestra el promedio de partidos sospechosos en cada superficie.

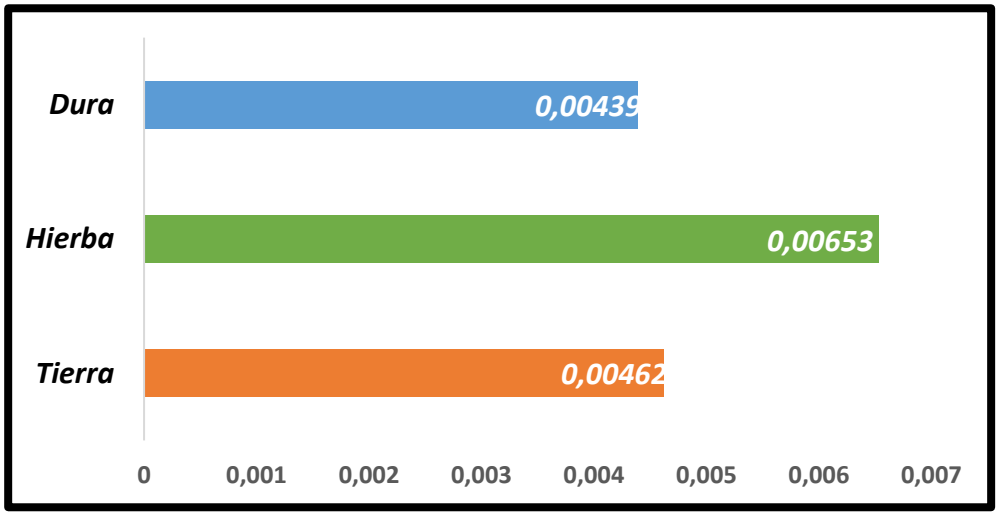


Figura 4.1: Porcentaje de partidos sospechosos por superficie.
Se muestra una gráfica de barras con los porcentajes de partidos sospechosos por cada superficie.

Partidos sospechosos por categoría:

- **En Grand Slam (ATP y WTA),** hay 135 (103 ATP y 32 WTA) partidos sospechosos, un 54.3% del total; significativo ya que sólo representan el 20.6% de los partidos.
- **En ATP500,** hay 12 partidos sospechosos, un 4.82% del total.
- **En ATP250**, hay 13 partidos sospechosos, un 5.22% del total.
- **En Masters 1000,** hay 18 partidos sospechosos, un 7.23% del total.
- **En Masters Cup,** no hay partidos sospechosos.
- **En International,** hay 46 partidos sospechosos, un 18.5% del total.
- **En Premier,** hay 25 partidos sospechosos, un 10% del total.
- **En Tour Championships,** no hay partidos sospechosos.

En la Figura 4.2 se muestra el promedio de partidos sospechosos en cada superficie.

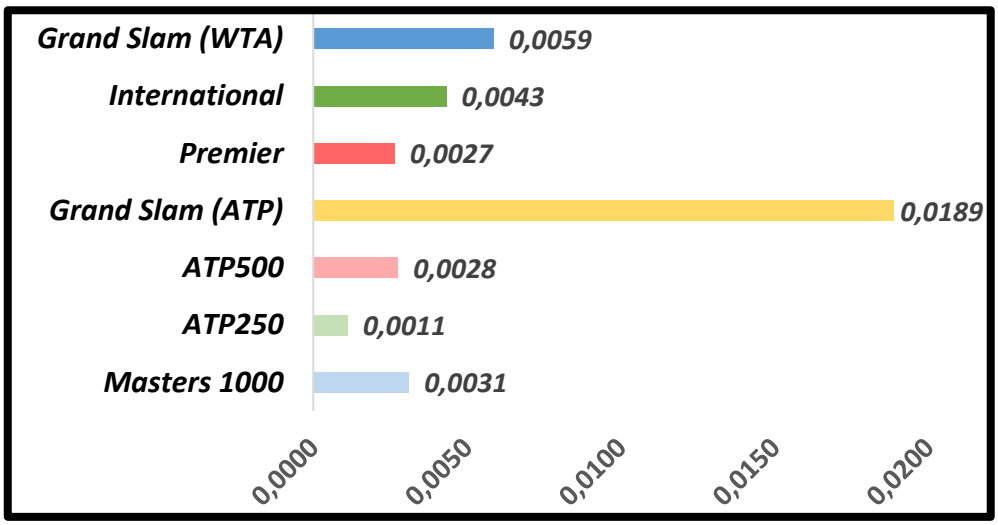


Figura 4.2: Gráfica de barras con el porcentaje de partidos sospechosos por categoría.
Se muestra una gráfica de barras con los porcentajes de partidos sospechosos por cada categoría.

5 CONCLUSIONES Y TRABAJO FUTURO

5.1 Conclusiones

En esta sección se enumeran las conclusiones obtenidas al final del proyecto, analizando si se cumplen los objetivos definidos:

- Se ha conseguido construir una base de datos con un gran nivel de información para hasta 53.054 partidos pertenecientes a los circuitos ATP y WTA.
- Se ha desarrollado un modelo de datos analítico para prevenir el fraude, basado en un modelo de *scoring* de riesgo que utiliza los indicadores definidos y en un modelo de detección de anomalías que utiliza el algoritmo Isolation Forest.
- La información de las redes sociales sobre los partidos señalados como sospechosos ha sido extraída y almacenada correctamente gracias al desarrollo de un programa que realiza peticiones a la API de Twitter y a su conexión la base de datos. En este punto se ha observado que es complicado encontrar comentarios en los que se comenten temas fraudulentos ya que este tipo de comentarios se suelen hacer en privado. Y, por otra parte, al extraer los comentarios de Twitter se ha observado como la proporción de tweets sobre partidos WTA es mucho menor a la de los partidos ATP con contadas excepciones como partidos de finales de Grand Slam o de las grandes estrellas del circuito femenino.
- Como ya se ha mencionado, se utiliza el algoritmo Isolation Forest para lograr el objetivo que definía el desarrollo de una herramienta de detección de partidos sospechosos de fraude. En cuanto a algoritmos de detección de anomalías, se puede concluir que el Isolation Forest es una de las soluciones más potentes; es eficaz y de sencilla implementación.
- En relación con el objetivo que definía la construcción de un *dashboard* en el que poder observar los datos resultantes del análisis realizado, se ha conseguido gracias a la tecnología Dash by Plotly que ha permitido desarrollar una gran herramienta de visualización. Es un sector que está en auge, de la mano del Big Data, y que está creciendo a pasos agigantados dando opción a crear aplicaciones web de análisis con una gran gama de posibilidades de personalización y con una gran facilidad de uso.
- Además de cumplir con todos los objetivos técnicos, como estaba previsto, el proceso también ha servido para aprender a utilizar nuevas tecnologías como la API de Twitter o Dash.

Para concluir, tras haber llevado a cabo el estudio de numerosos partidos de tenis y haber leído numerosos comentarios y artículos, es una evidencia que el fraude de partidos ha existido y sigue existiendo en este deporte; y que es uno de los principales problemas que debe tratar de solucionar tanto el tenis como muchos otros deportes.

5.2 Trabajo futuro

El trabajo realizado en este proyecto se puede utilizar de base para un proyecto mayor, como se describe a continuación:

Por una parte, se podría ampliar el estudio a un mayor número de competiciones de tenis e incluso de otros deportes como el pádel sin necesidad de demasiados cambios en el sistema al ser deportes con número de jugadores, puntajes y marcadores similares.

Por otra parte, se podría llevar a cabo un análisis más amplio en el caso de poder obtener datos privados de las casas de apuestas o un acceso ilimitado a comentarios en Twitter y otras redes sociales; lo que probablemente daría lugar a un mejor resultado a la hora de detectar las anomalías.

Además, contando con estos datos también se podría crear un modelo de *scoring* más certero, fundamentado en un mayor número de datos.

Por último, el análisis también podría crecer a nivel de jugador, ya que este trabajo está enfocado a nivel de evento; y teniendo más información sobre los jugadores se podrían crear visualizaciones interesantes en la aplicación web.

REFERENCIAS

[1] "Anuario del juego en España 2020", *UC3M y CEJUEGO*, 2020 [En línea]. Disponible en: https://cejuego.com/wp-content/uploads/2021/04/Anuario_del_Juego_en_Espan%CC%83a_2020.pdf. [Accedido: 1-jun-2021].

[2] España, "Ley 13/2011, de 27 de mayo, de regulación del juego.", *Boletín Oficial del Estado,* 28 de mayo de 2011, núm. 127, Artículo 3 y Artículo 6.

[3] "Apuesta deportiva", *Wikipedia,* 2021. [En línea]. Disponible en: https://es.wikipedia.org/wiki/Apuesta_deportiva. [Accedido: 1-jun-2021].

[4] I. López Gimeno, "Historia de las Apuestas en España", *ApuestasOnline.net*, 2020. [En línea]. Disponible en: https://apuestasonline.net/historia-apuestas-espana/. [Accedido: 1-jun-2021].

[5] "Juego online en España, datos y estadísticas", *Europa Press*, 2020. [En línea]. Disponible en: https://www.epdata.es/datos/juego-online-espana-datos-estadisticas/161. [Accedido: 1-jun-2021].

[6] J. Lacort, ""Las apuestas son la heroína del siglo XXI": así es como las casas apuestas online han conquistado España", *Xataka*, 2020. [En línea]. Disponible en: https://www.xataka.com/especiales/apuestas-heroina-siglo-xii-asi-como-apuestas-online-han-conquistado-espana. [Accedido: 1-jun-2021].

[7] A. D. Prieto, "El Gobierno prohíbe por decreto la publicidad de las casas de apuestas por suponer 'alarma social'", *El Español*, 2020. [En línea]. Disponible en: https://www.elespanol.com/espana/politica/20201103/gobierno-prohibe-publicidad-apuestas-considera-alarma-social/533197277_0.html. [Accedido: 1-jun-2021].

[8] H. Blake y J. Templon, "The Tennis Racket", *Buzz Feed News*, 2016. [En línea]. Disponible en: https://www.buzzfeednews.com/article/heidiblake/the-tennis-racket. [Accedido: 1-jun-2021].

[9] R. Paniagua, "La lacra de las apuestas y las mafias: 'Si ganas estás muerto'", *El Periódico*, 2017. [En línea]. Disponible en: https://www.elperiodico.com/es/deportes/20170411/apuestas-mafias-deporte-ganas-estas-muerto-5958435. [Accedido: 1-jun-2021].

[10] J. F. Díaz, "Los amaños de partidos también existen en los torneos del Grand Slam", *El Confidencial*, 2013. [En línea]. Disponible en: https://www.elconfidencial.com/deportes/tenis/2013-12-12/los-amanos-de-partidos-tambien-existen-en-los-torneos-del-grand-slam_65326/. [Accedido: 1-jun-2021].

[11] P. Pareja, "El oscuro mundo de los 'tipsters', los pronosticadores que ejercen de gancho de las casas de apuestas", *El Diario*, 2020. [En línea]. Disponible en: https://www.eldiario.es/catalunya/oscuro-tipsters-pronosticadores-apuestas-deportivas_1_1081705.html. [Accedido: 1-jun-2021].

[12] R. Varea, "La herramienta antiamaños más potente del mundo", *El País*, 2020. [En línea]. Disponible en: https://elpais.com/economia/2020/04/19/actualidad/1587289180_012133.html. [Accedido: 1-jun-2021].

[13] Premium Search Tweets: Full-Archive API|Docs|Twitter Developer, 2021. [En línea]. Disponible en: https://developer.twitter.com/en/docs/twitter-api/premium/search-api/quick-start/premium-full-archive. [Accedido: 1-jun-2021].

[14] A. Carreño López, "Detección de sucesos raros con machine learning", TFM, Universidad Politécnica de Madrid, 2017 [En línea]. Disponible en: http://oa.upm.es/47931/. [Accedido: 1-Jun-2021].

[15] M. Barus, "Anomaly Detection with Isolation Forests using H20", *H20.ai*, 2018. [En línea]. Disponible en: https://www.h2o.ai/blog/anomaly-detection-with-isolation-forests-using-h2o/. [Accedido: 1-Jun-2021].

[16] R. Caruana y A. Niculescu-Mizil. "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning - ICML '06, vol. 2006, pp. 161-168., June 2011.

[17] S.Ding, Z. Zhu y X. Zhang. "An overview on semi-supervised support vector machine". *Neural Computing and Applications*, vol. 28, pp. 969-978, May 2017.

[18] J. Bella Santos, "Métodos de aprendizaje automático para detección de anomalías", TFM, Universidad Autónoma de Madrid, 2019 [En línea]. Disponible en: http://hdl.handle.net/10486/688762. [Accedido: 1-Jun-2021].

[19] J. Amat Rodrigo, "Detección de anomalías: Isolation Forest", Ciencia de datos, 2020. [En línea]. Disponible en: https://www.cienciadedatos.net/documentos/66_deteccion_anomalias_isolationforest.html. [Accedido: 1-Jun-2021].

[20] N. Forteza, "Isolation Forest: el algoritmo estrella para detección de anomalías", *keepler.io*, 2019. [En línea]. Disponible en: https://medium.com/@keepler.io/isolation-forest-el-algoritmo-estrella-para-detecci%C3%B3n-de-anomal%C3%ADas-416bb5892f10. [Accedido: 1-Jun-2021].

[21] Scikit-learn 0.24.2 documentation, 2021. [En línea]. Disponible en: https://scikit-learn.org/stable/. [Accedido: 1-jun-2021].

[22] Dash Documentation & User Guide|Plotly, 2021. [En línea]. Disponible en: https://dash.plotly.com/. [Accedido: 1-jun-2021].

[23] Tennis-Data.co.uk, 2021. [En línea]. Disponible en: http://tennis-data.co.uk/alldata.php. [Accedido: 1-jun-2021].

[24] Bet365, 2021. [En línea]. Disponible en: https://www.bet365.es. [Accedido: 1-jun-2021].

[25] Pinnacle, 2021. [En línea]. Disponible en: https://www.pinnacle.com/es/. [Accedido: 1-jun-2021].

[26] Psycopg 2.8.7.dev0 documentation, 2021. [En línea]. Disponible en: https://www.psycopg.org/docs/. [Accedido: 1-jun-2021].

[27] Pandas documentation, 2021. [En línea]. Disponible en: https://pandas.pydata.org/pandas-docs/stable/index.html. [Accedido: 1-jun-2021].

[28] ISO – ISO 3166 – Country Codes, 2021. [En línea]. Disponible en: https://www.iso.org/iso-3166-country-codes.html. [Accedido: 1-jun-2021].

[29] Quickstart – dbc docs, 2021. [En línea]. Disponible en: https://dash-bootstrap-components.opensource.faculty.ai/docs/. [Accedido: 1-jun-2021].

Anexos

A| Visualización del sistema

En este anexo se muestran imágenes del visualizador creado en el proyecto.

La navegación dentro de la aplicación se realiza a través de las pestañas que se encuentran en la parte superior.

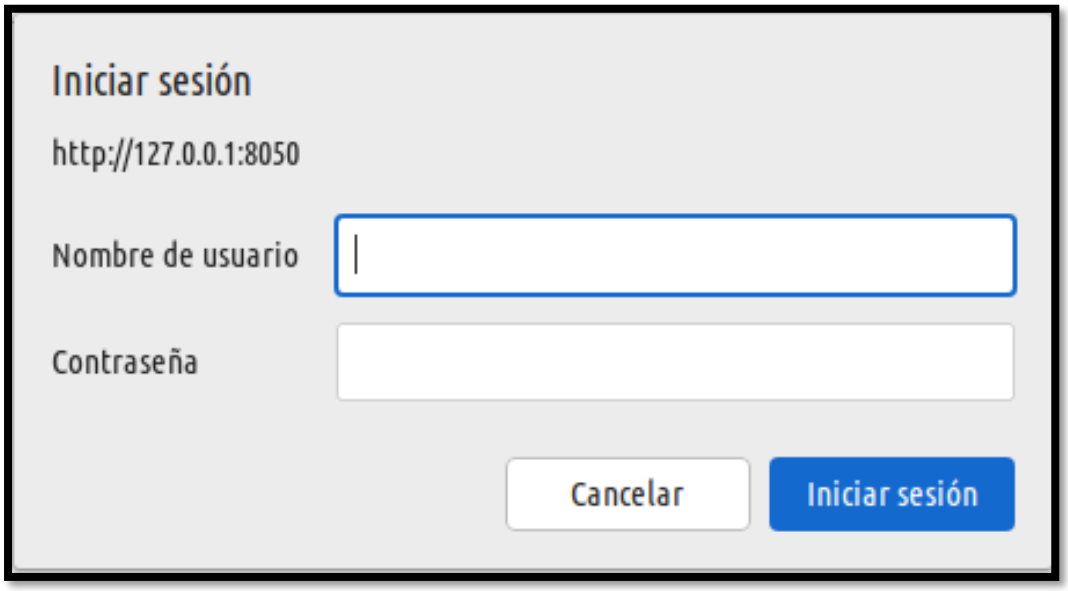


Figura A.1: Login.
Pantalla de inicio de sesión del visualizador

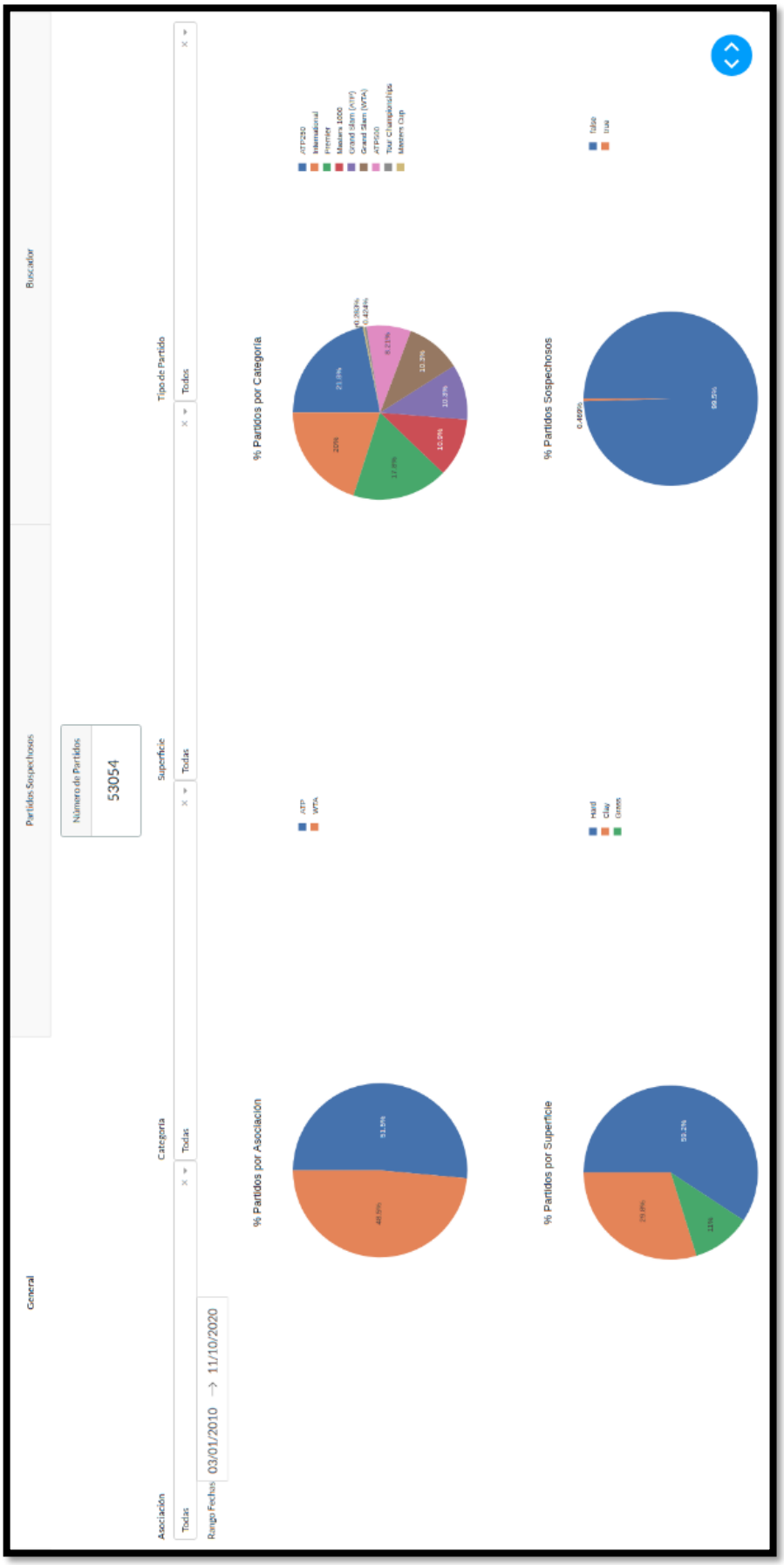


Figura A.2: Filtros y gráficas de tarta.
Pestaña general del visualizador, se muestran los filtros y las gráficas de tarta disponibles.

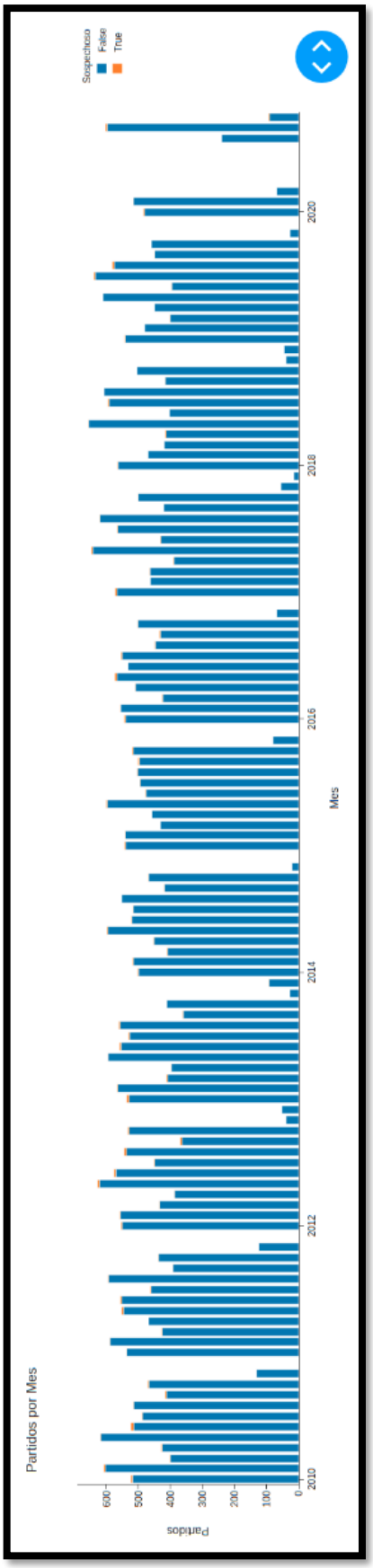


Figura A.3: Gráfica de barras con los partidos por mes.
Pestaña general del visualizador, se muestra la gráfica de barras disponible.

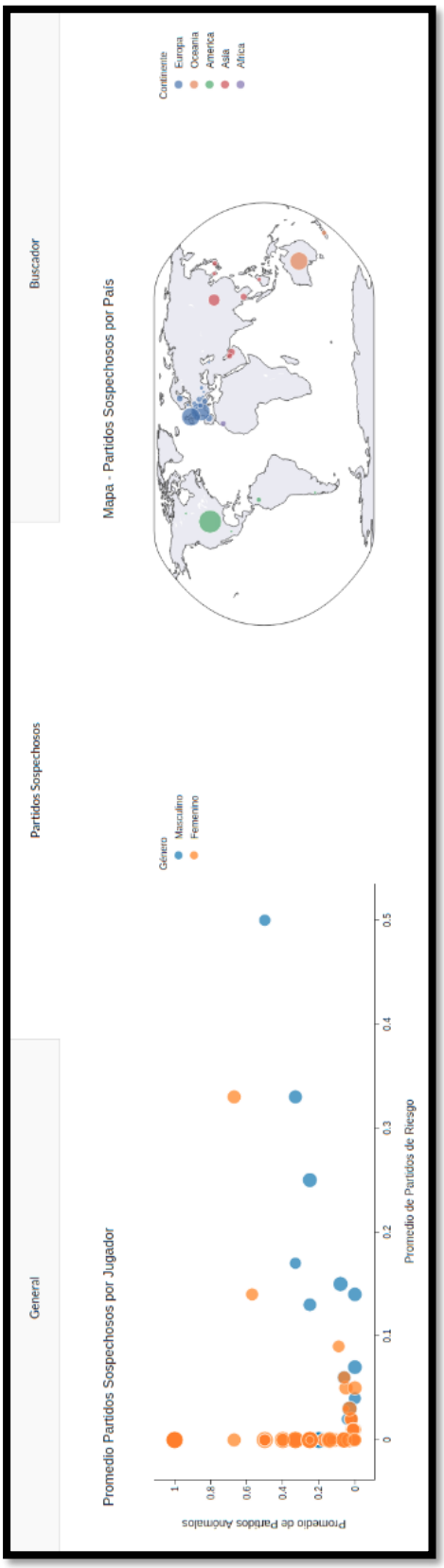


Figura A.4: Gráfica de burbujas y mapa con los partidos sospechosos.
Pestaña de partidos sospechosos del visualizador, se muestran la gráfica de burbujas y el mapa disponibles.

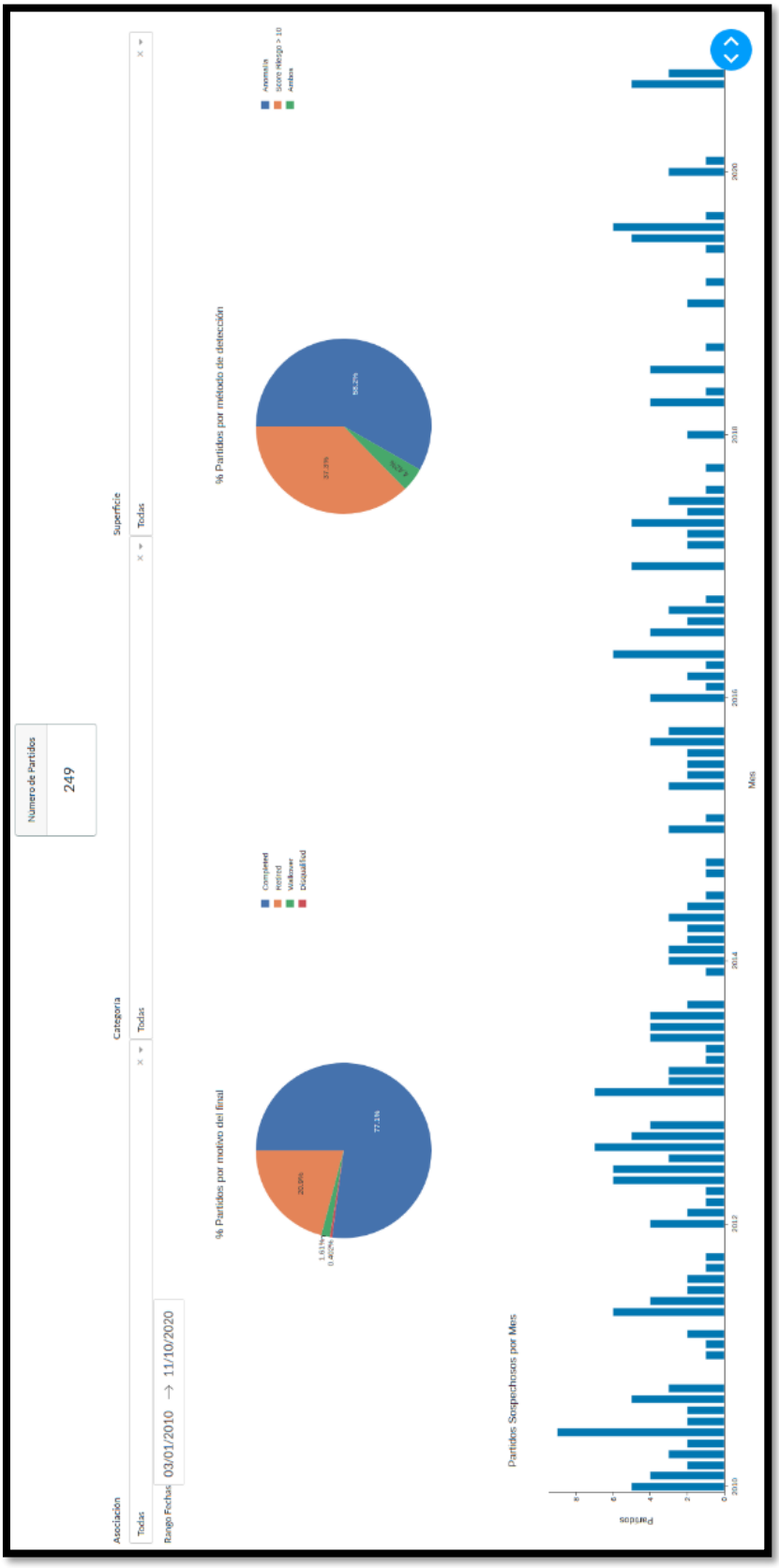


Figura A.5: Gráficas de tartas y gráfica de barras de partidos sospechosos. Pestaña de partidos sospechosos del visualizador, se muestran los filtros, las gráficas de tarta y las gráficas de barras disponibles.

Figura A.6: Información de jugador.
Pestaña del buscador del visualizador, se muestra información sobre el jugador buscado (datos anonimizados).

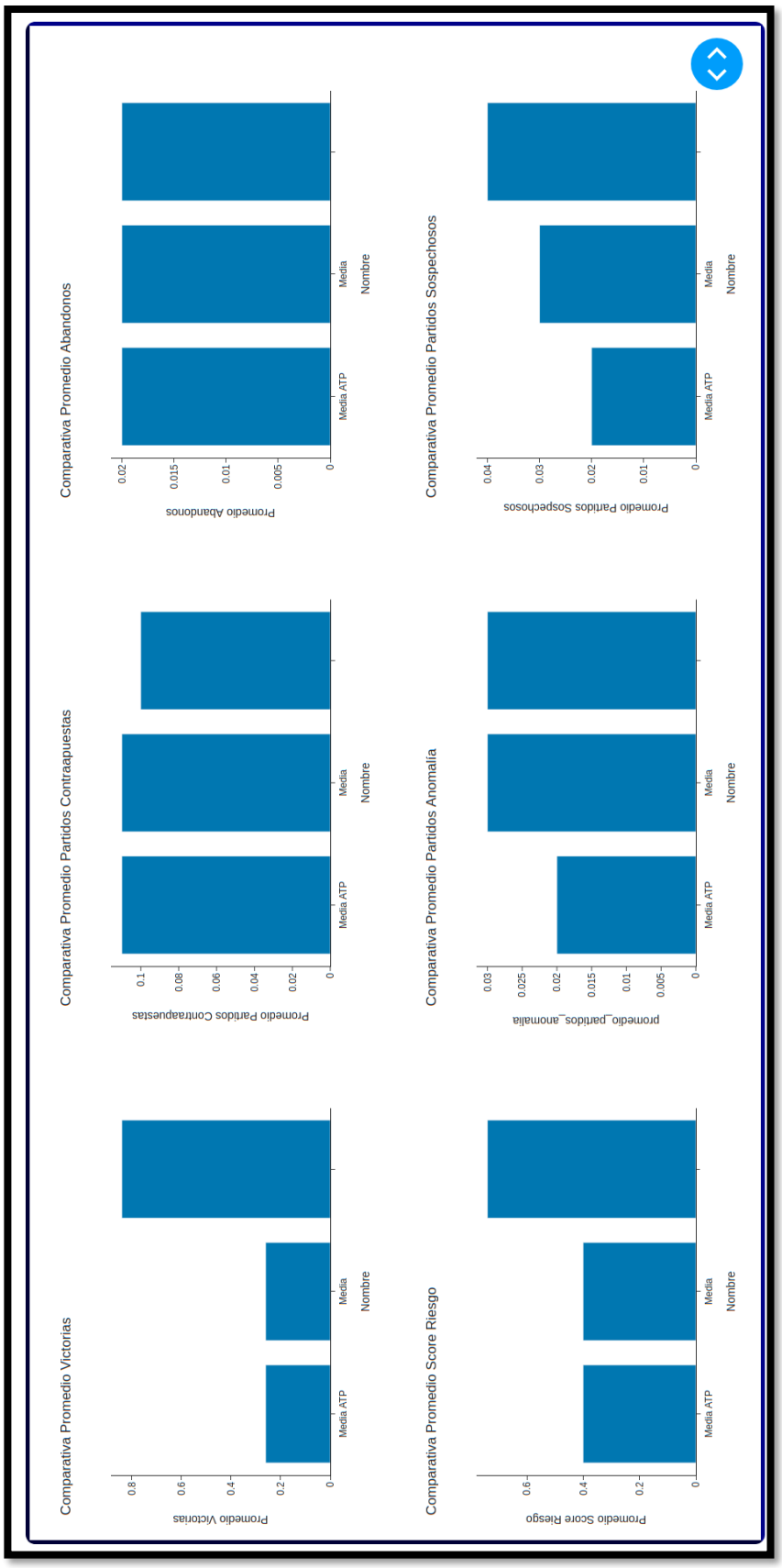


Figura A.7: Gráficas de barras comparativas de jugador.
Pestaña del buscador del visualizador, se muestran comparativas del jugador buscado.

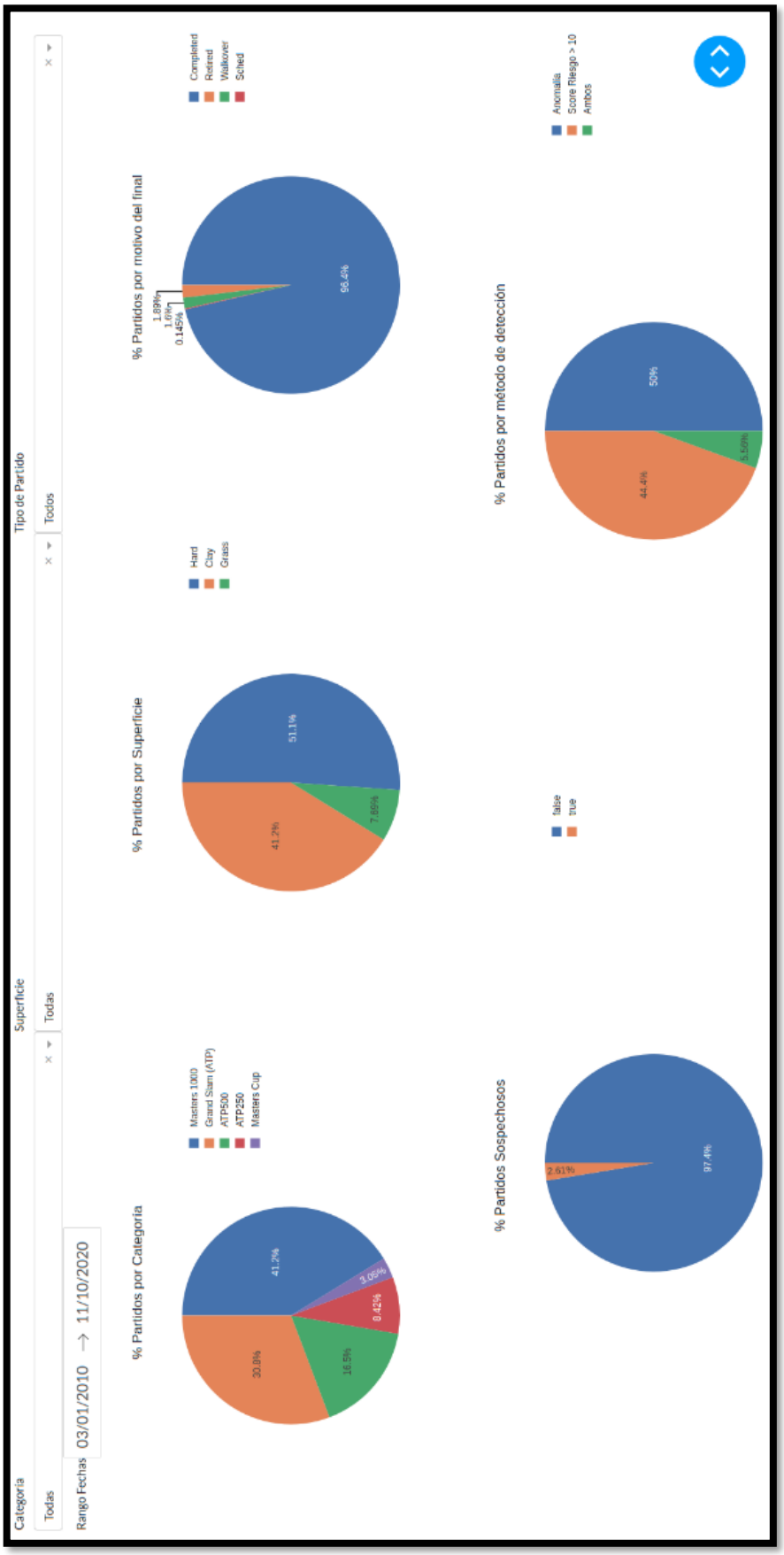


Figura A.8: Filtros y gráficas de tarta de jugador.
Pestaña del buscador del visualizador, se muestran los filtros y las gráficas de tarta disponibles sobre el jugador buscado.

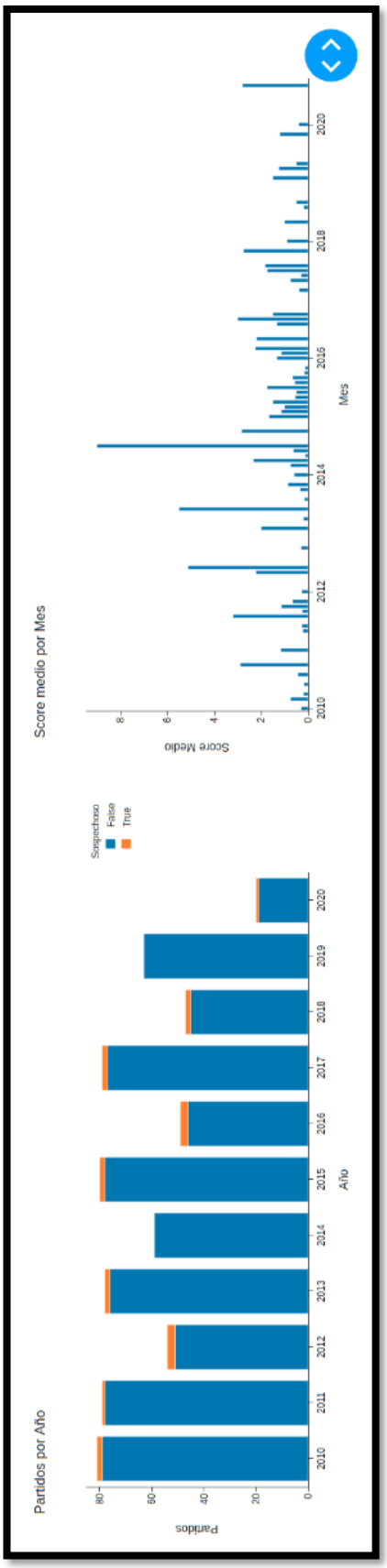


Figura A.9: Partidos por año y score medio por mes del jugador.
Pestaña del buscador del visualizador, se muestran las gráficas de barras disponibles sobre el jugador buscado.

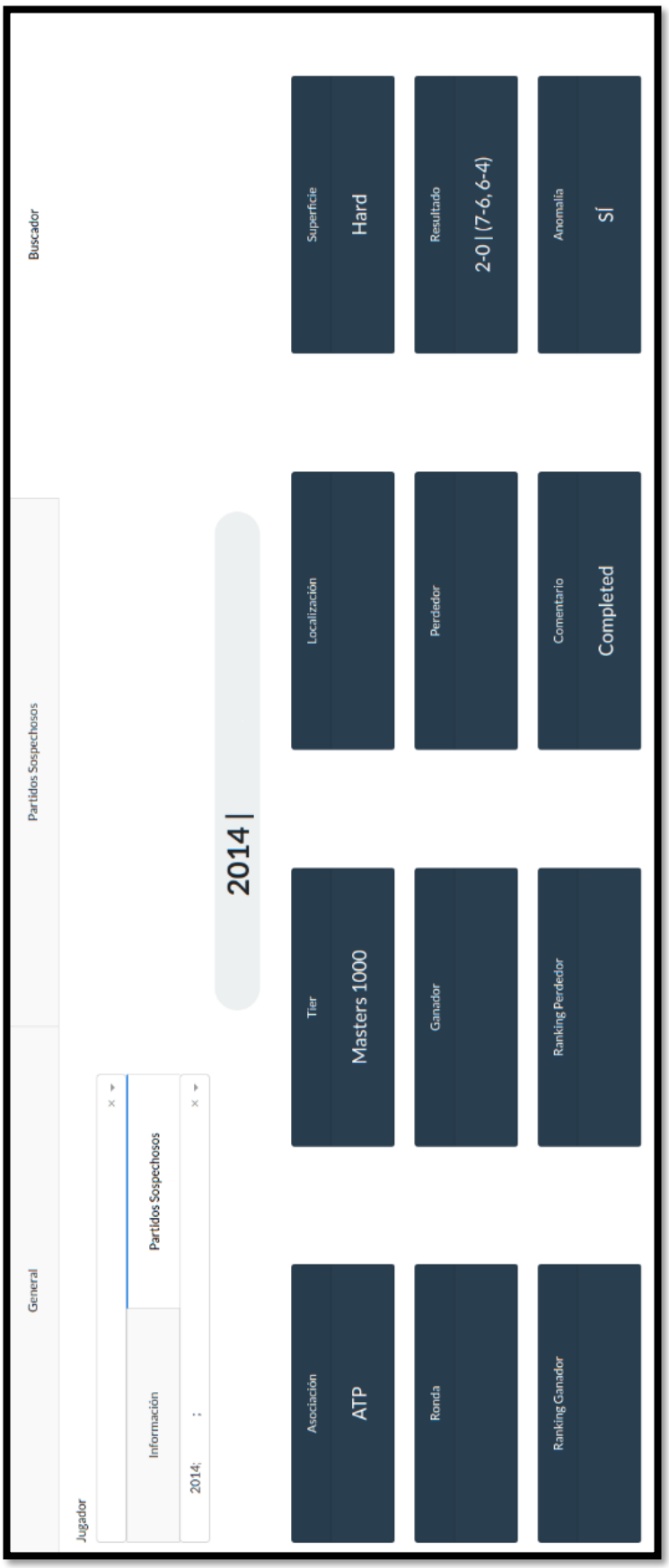


Figura A.10: Tarjetas con la información del partido.
Pestaña del buscador del visualizador, se muestra información sobre el partido buscado (datos anonimizados).

Figura A.11: Tarjetas comparativas de información del partido.
Pestaña del buscador del visualizador, se muestran comparativas del partido buscado.

Tweets sobre el partido

Fecha		Comentario	Retweets	Favs	Respuestas	Usuario	Seguidores	Siguiendo	Verificado
11:18 - 07/10/2014	@	Estaba amañando, se pagaba a 11 euros, no es la 1era de mi sera la ultima, troll de trolls, http://t.co/XDDkqhtmiq	0	0	1		124	133	No
08:50 - 07/10/2014		Si señores, lo ha vuelto a hacer, pierde contra el del mundo. A 13 pavetes se pagaba la victoria de ;	0	0	0		1311	0	No
08:36 - 07/10/2014		Ojo a la proeza que puede conseguir , perder ante del mundo, va set y break abajo #tenis #apuestas	0	0	0		3075	222	
18:22 - 06/10/2014		1.04, 9.00 en @bet365_es ¿#Apuestas en el #ShanghaiMasters? http://t.co/eK2S6jvNbx #Tenis	0	0	0		682	748	

Figura A.12: Comentarios extraídos de Twitter sobre el evento.
Pestaña del buscador del visualizador, se muestra los *tweets* sobre el partido buscado (datos anonimizados).

B| Thinking aloud

En este anexo se incluyen las tareas y el formulario que forman parte de la técnica 'Thinking Aloud'.

Para conocer la usabilidad del visualizador de datos desarrollado, se lleva a cabo la técnica 'Thinking Aloud' que consiste en proponer a varios usuarios expresar sus pensamientos y opiniones mientras realiza un recorrido por las distintas funcionalidades del dashboard a través de la realización de distintas tareas que están descritas a continuación y que se estiman en una duración de 15 minutos. Al final de estas tareas, también se propone realizar un formulario que completará los comentarios recibidos durante el *test*.

1) **Acceder a la herramienta**
   a) Introducir en el navegador la dirección donde se encuentra disponible la herramienta.
   b) Introducir el nombre de usuario y la contraseña dados.

2) **Eres un usuario de la aplicación que quiere evaluar cuantos partidos sospechosos de fraude se han detectado en los últimos cinco años en las competiciones de Grand Slam ATP disputadas en tierra.**
   a) La primera tarea a realizar una vez dentro del visualizador será observar las cinco gráficas de la pestaña general y filtrar por:
      • Asociación: ATP.
      • Categoría: Grand Slam (ATP)
      • Superficie: Clay
      • Rango Fechas: 01/01/2015 -> 11/10/2020
   b) Observar la información relevante de las gráficas '% Partidos Sospechosos' y 'Partidos por Mes' utilizando el cursor.
   c) Descargar una imagen de la gráfica de barras en formato png.

3) **Ahora como usuario de la aplicación quieres analizar los partidos sospechosos en más profundidad y conocer más información sobre los partidos anteriormente filtrados.**
   a) Completada la tarea anterior y descargada la gráfica, cambiar a la pestaña de partidos sospechosos.
   b) Observar en el mapa en que países se han detectado más partidos sospechosos, utilizar el cursor para obtener más información de la que aparece a simple vista.
   c) Realizar el mismo filtrado que en la primera tarea, y utilizar de nuevo el cursor para poder ver más datos sobre las gráficas disponibles.
   d) Obtener la lista de partidos detectados como sospechosos para cualquier mes como puede ser junio de 2019, recordar esta información obtenida que se utilizará en siguientes tareas.

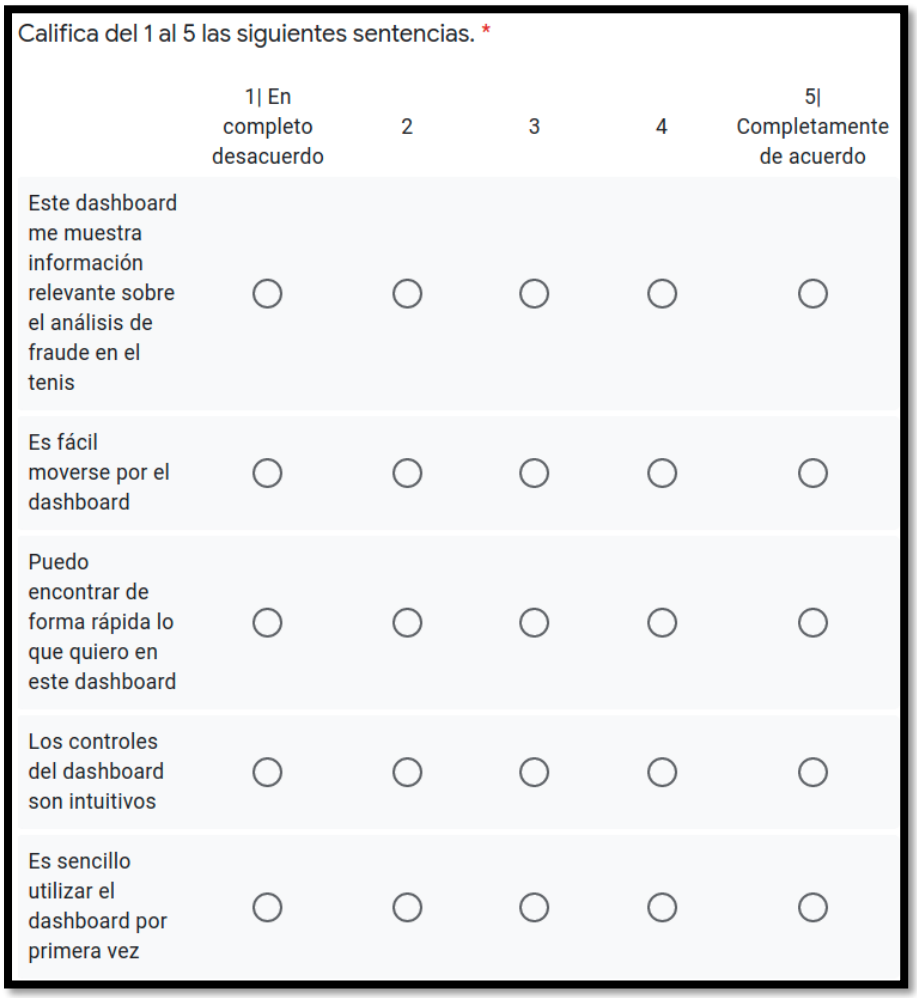
4) **Como usuario de la aplicación una vez que has conocido un partido detectado como sospechoso por la aplicación, tienes intención de profundizar en el jugador que ha sido derrotado en ese partido.**
   a) Cambiar a la pestaña del buscador.
   b) Buscar al jugador del que se quiere obtener más datos.
   c) Interactuar con las distintas gráficas disponibles.

d) Realizar un filtrado seleccionando la superficie ‘Clay’ y como fecha inicial el 01/01/2015.
e) Interactuar con la gráfica de ‘Score medio por Mes’ y ver el score de riesgo medio del jugador durante los meses filtrados.
f) En la gráfica contigua, mostrar sólo los partidos sospechosos en la gráfica ‘Partidos por Año’ y obtener el listado de partidos sospechosos para cualquiera de las barras que persisten en la gráfica.

5) Finalmente, para terminar de informarnos sobre el partido en el cual se ha puesto la mira, se accederá a toda la información que hay sobre este.
a) Acceder dentro del buscador a la pestaña de partidos sospechosos.
b) Buscar el partido del que se quieren conocer más datos.
c) Analizar la información, las comparativas y los tweets relacionados en caso de que existan para dicho partido.

6) Se abandona la aplicación web y se accede al siguiente formulario que complementará los comentarios dados durante la realización de las tareas.

Se propone responder al formulario mostrado en la Figura B.1 y en la Figura B.2:



Califica del 1 al 5 las siguientes sentencias. *

	1| En completo desacuerdo	2	3	4	5| Completamente de acuerdo
Este dashboard me muestra información relevante sobre el análisis de fraude en el tenis	○	○	○	○	○
Es fácil moverse por el dashboard	○	○	○	○	○
Puedo encontrar de forma rápida lo que quiero en este dashboard	○	○	○	○	○
Los controles del dashboard son intuitivos	○	○	○	○	○
Es sencillo utilizar el dashboard por primera vez	○	○	○	○	○

Figura B.1: Cinco primeras sentencias del formulario.
Se muestran las cinco primeras sentencias evaluables del formulario.

Figura B.2: Tres últimas sentencias del formulario.
Se muestran las tres últimas sentencias evaluables del formulario.

Por último, se encuentran las siguientes cuestiones:

- ¿Cuáles consideras que son las mejores características de este *dashboard*, y por qué?
- ¿Qué característica/características del *dashboard* consideras que deberían mejorarse, y por qué?
- ¿Cuánto tiempo has tardado en completar las tareas?
- ¿Te ha resultado fácil completar las tareas?
- En caso negativo, ¿Por qué?
- ¿Cuántas tareas has completado?
- Comentarios adicionales